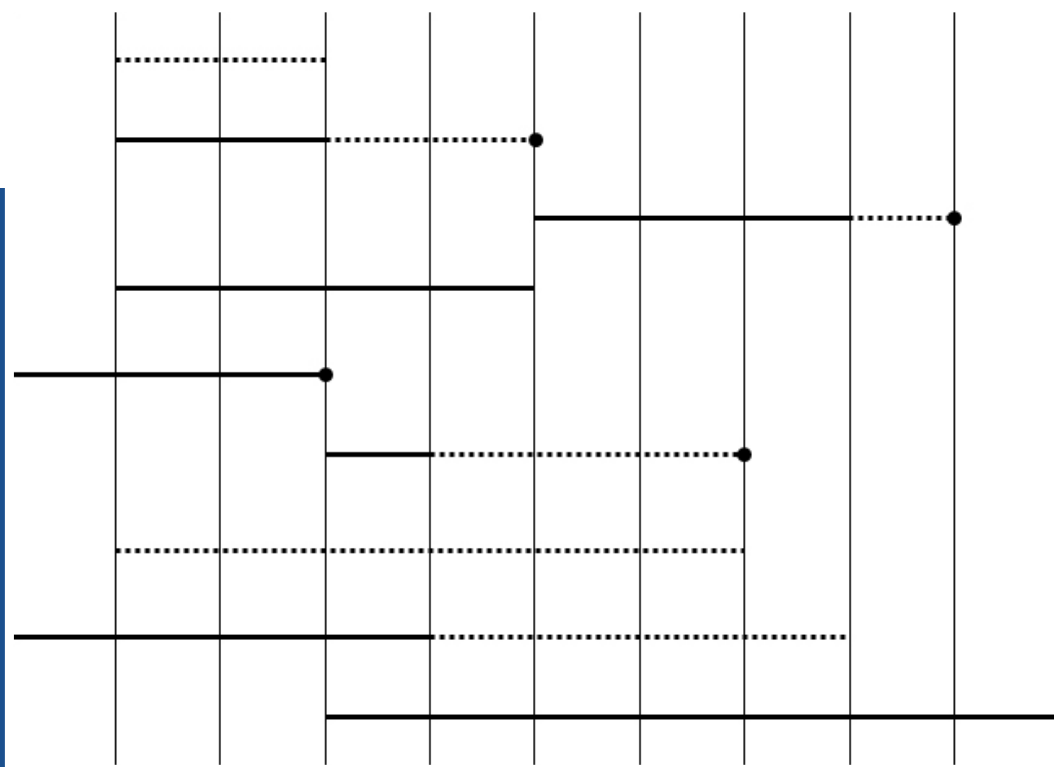


Epidemiología básica



GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES



**Escuela Nacional de Sanidad
Instituto de Salud Carlos III
Ministerio de Ciencia, Innovación y Universidades**

Avda. Monforte de Lemos 5
28029 MADRID (ESPAÑA)

Publicación incluida en el programa editorial del Ministerio de Ciencia, Innovación y Universidades.

Catálogo general de publicaciones oficiales:

<https://cpage.mpr.gob.es/>

Para obtener este informe de forma gratuita en Internet :

<http://hdl.handle.net/20.500.12105/17538>



<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es>

Edita:

**INSTITUTO DE SALUD CARLOS III.
Ministerio de Ciencia, Innovación y Universidades**

NIPO pdf: 156240125

NIPO: 15624011X

D.L.: M-5162-2024

DOI: [10.4321/repisalud.17538](https://doi.org/10.4321/repisalud.17538)

Diseño y maquetación: Diseño Gráfico Gallego y Asociados, S.L.

Editores:

Miguel Ángel Royo Bordonada. Escuela Nacional de Sanidad, Instituto de Salud Carlos III.

Javier Damián. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Fernando José García López. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Autores:

Nuria Aragonés Sanz. Dirección General de Salud Pública, Consejería de Sanidad de Madrid.

José Ramón Banegas Banegas. Departamento de Medicina Preventiva y Salud Pública, Universidad Autónoma de Madrid.

Gregorio Barrio Anta. Escuela Nacional de Sanidad, Instituto de Salud Carlos III.

María José Belza Egozcue. Escuela Nacional de Sanidad, Instituto de Salud Carlos III.

Jesús Castilla Catalán. Instituto de Salud Pública de Navarra.

Ferrán Catalá López. Escuela Nacional de Sanidad, Instituto de Salud Carlos III.

Isabel del Cura González. Gerencia Asistencial de Atención Primaria, Servicio Madrileño de Salud.

Laura del Nido. Hospital Universitario General de Villalba.

Javier Damián. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Iñaki Galán Labaca. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Esther Elena García Carpintero. Agencia de Evaluación de Tecnologías Sanitarias, Instituto de Salud Carlos III.

Fernando José García López. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Jesús González Enríquez. Agencia de Evaluación de Tecnologías Sanitarias, Instituto de Salud Carlos III.

Iñaki Imaz Iglesia. Agencia de Evaluación de Tecnologías Sanitarias, Instituto de Salud Carlos III.

Jesús Martín Fernández. Gerencia Asistencial de Atención Primaria, Servicio Madrileño de Salud.

Beatriz Pérez Gómez. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Elena Plans Beriso. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Marina Pollán Santamaría. Centro Nacional de Epidemiología, Instituto de Salud Carlos III.

Fernando Rodríguez Artalejo. Departamento de Medicina Preventiva y Salud Pública, Universidad Autónoma de Madrid.

Miguel Ángel Royo Bordonada. Escuela Nacional de Sanidad, Instituto de Salud Carlos III.

Belén Sanz Barbero. Escuela Nacional de Sanidad, Instituto de Salud Carlos III.

Para citar esta monografía:

Royo Bordonada MÁ, Damián J, García López FJ. Epidemiología básica. Madrid: Instituto de Salud Carlos III; 2024.

Índice

Capítulo 1. Conceptos, usos e historia de la epidemiología	9
Capítulo 2. Medidas de frecuencia y de efecto	47
Capítulo 3. Tipos de estudios epidemiológicos	75
Capítulo 4. Sesgos	97
Capítulo 5. Confusión e interacción.....	133
Capítulo 6. Análisis de datos epidemiológicos	151
Capítulo 7. Gráficos causales.....	181
Capítulo 8. Encuestas de salud poblacionales.....	205
Capítulo 9. Estudios de intervención	233
Capítulo 10. Estudios de cohortes.....	285
Capítulo 11. Estudios de casos y controles	321
Capítulo 12. Ensayos por conglomerados	343
Capítulo 13. Estudios diagnósticos. Validez y fiabilidad.....	369
Capítulo 14. El cribado en la salud pública.....	391
Capítulo 15. Revisiones sistemáticas y metanálisis	417
Capítulo 16. Planificación y ejecución de un cuestionario	455
Capítulo 17. Inferencia causal en epidemiología	483
Glosario de términos básicos en epidemiología	523

Prefacio

Los primeros antecedentes de la epidemiología datan de la época de Hipócrates (siglo IV a.C.); pero no fue hasta el siglo XVII cuando se empezaron a realizar, de forma aislada, algunos estudios de investigación de enfermedades siguiendo un método científico que sentaría las bases del método epidemiológico. Los estudios de John Graunt (1620-1674), que construyó las primeras tablas de vida (tablas de mortalidad), James Lind (1716-1794), que llevó a cabo el primer ensayo clínico de la historia en enfermos de escorbuto, Edward Jenner (1749-1823), que dio con la primera vacuna de la viruela, y John Snow (1813-1858), que demostró que el cólera se transmitía por contaminación fecal del agua de consumo, son algunos de los más carismáticos entre los precursores de la epidemiología moderna. Pero no fue hasta mediados del siglo XX cuando la epidemiología empezó a consolidarse como una disciplina científica independiente y con identidad propia. Algunos de los estudios epidemiológicos que han tenido mayor trascendencia pública datan de esa época. Por ejemplo, las investigaciones de Doll y Hill sobre el efecto del tabaco como agente causal del cáncer de pulmón, los estudios de Keys y Grande Covián sobre la relación entre las grasas de la dieta y los lípidos plasmáticos, que sentaron las bases de la moderna epidemiología nutricional, los ensayos de intervención comunitaria para evaluar la eficacia de los suplementos de flúor a través de la red de agua de consumo en la prevención de la caries dental o el *Framingham Study*, un estudio de seguimiento diseñado por el Servicio de Salud Pública de los Estados Unidos en la década de los 40 para investigar los factores implicados en la cardiopatía isquémica que, todavía hoy, sigue proporcionando importantes resultados.

El diccionario de epidemiología define esta disciplina esencial de la salud pública como «el estudio de la aparición y distribución de sucesos, estados y procesos relacionados con la salud en poblaciones específicas, incluyendo el estudio de los determinantes que influyen en esos procesos y la aplicación de este conocimiento para controlar los problemas de salud relevantes». La Escuela Nacional de Sanidad es una institución centenaria de investigación y formación de posgrado en salud pública mediante diversas acciones formativas, entre las que destacan cursos especializados de formación continua, diplomas, como los de salud pública o cooperación internacional, y másteres, como los de administración sanitaria o salud pública. Desde sus orígenes, la Escuela Nacional de Sanidad ha tenido entre sus objetivos la elaboración de materiales docentes para sus cursos. Con esta idea nació la primera edición del *Manual de método epidemiológico*, que contenía 13 capítulos sobre los conceptos y métodos esenciales del método epidemiológico, elaborados por los profesionales que

en esos años impartían la docencia de esos temas en la Escuela Nacional de Sanidad. Esta primera publicación tuvo una difusión amplia pero limitada, ya que se editó únicamente en papel, y en 2009 se realizó una edición actualizada en formato electrónico, de acceso libre (<http://hdl.handle.net/20.500.12105/5271>). La disciplina ha evolucionado mucho en los últimos años, haciendo necesaria la incorporación a la docencia de nuevos conceptos y métodos, lo que ha motivado la elaboración de la presente edición. Este volumen se estructura en 17 capítulos desarrollados por 21 profesores e investigadores con larga experiencia docente en epidemiología. Se mantienen la mayoría de los capítulos de ediciones anteriores, si bien con las actualizaciones convenientes y, en algunos casos, con ampliación de sus contenidos, como en los capítulos de casos y controles, estudios de intervención e inferencia causal. El tema de validez se ha dividido en 2 capítulos, sesgos por una parte y confusión e interacción por otra. Y se incluyen un glosario y 5 capítulos nuevos: gráficos causales, encuestas poblacionales, ensayos por conglomerados, cribado, y planificación y ejecución de un cuestionario.

Este libro constituye una herramienta básica de apoyo a la docencia en epidemiología, pero conviene tener presente la conveniencia de consultar otros textos para profundizar en los conceptos y métodos epidemiológicos descritos a lo largo de sus capítulos. Esperamos que también contribuya a estimular el estudio y la práctica rigurosa del método epidemiológico para su aplicación en distintos ámbitos, desde la vigilancia de salud pública y la investigación causal hasta la práctica clínica o la investigación de servicios de salud, entre otros campos donde esta disciplina científica resultará de gran utilidad.

MIGUEL ÁNGEL ROYO, JAVIER DAMIÁN y FERNANDO GARCÍA LÓPEZ

Capítulo 1.

Conceptos, usos e historia de la epidemiología

FERNANDO GARCÍA LÓPEZ

1. Concepto y usos de la epidemiología

- 1.1. Definición de epidemiología
- 1.2. Objetivos específicos de la epidemiología
- 1.3. Conceptos de la epidemiología
- 1.4. Epidemiología y prevención
- 1.5. La epidemiología y la práctica clínica
- 1.6. El enfoque epidemiológico

2. Historia de la epidemiología

- 2.1. Epidemiología incipiente
- 2.2. Epidemiología temprana
- 2.3. Epidemiología clásica
- 2.4. Epidemiología moderna

3. Referencias bibliográficas

1. Concepto y usos de la epidemiología

1.1. Definición de epidemiología

Según *A Dictionary of Epidemiology*, (1) «la epidemiología es el estudio de la aparición y distribución de los acontecimientos, situaciones y procesos relacionados con la salud en poblaciones específicas, lo que incluye el estudio de los determinantes que influyen en esos procesos, y la aplicación de este conocimiento para el control de problemas de salud relevantes».

Veamos con detalle los distintos componentes de esta definición:

- *el estudio* puede ser la vigilancia, la observación, el cribado, la comprobación de hipótesis, la investigación analítica de relaciones causales, la experimentación y la predicción;
- *los acontecimientos, situaciones y procesos relacionados con la salud* abarcan fenómenos tan dispares como los brotes epidémicos, cualquier enfermedad o trastorno, las causas de muerte, las conductas, los procesos ambientales o socioeconómicos que afectan a la salud, la calidad de vida relacionada con la salud o la provisión y utilización de los servicios sanitarios;
- *las poblaciones específicas* son aquellas con características identificables comunes. Por ejemplo, los habitantes de la ciudad de Toledo en un determinado año o las mujeres españolas usuarias de anticonceptivos orales;¹
- *la aparición y distribución* tienen que ver con el análisis de los interrogantes de cuándo, dónde y quiénes —es decir, el tiempo, el lugar y la persona— de los fenómenos relacionados con la salud;
- *los determinantes* se refieren a los factores genéticos, geofísicos, biológicos, conductuales, nutricionales, sociales, culturales, económicos, políticos, ambientales y de otro tipo que influyen en la salud, y
- *la aplicación de este conocimiento para el control de problemas de salud relevantes* quiere decir promover, proteger y restaurar la salud, así como generar conocimiento científico.²

¹ La epidemiología, además de estudiar las poblaciones humanas, también se aplica a poblaciones de animales. En sentido estricto, el equivalente a las epidemias en los seres humanos en otras especies animales son las epizootias. El estudio de las epizootias se denomina epizootiología, epizootología o, también, epidemiología veterinaria.

² La etimología de epidemiología no es muy esclarecedora. La palabra *epidemia*, que aparece registrada en la lengua española al menos desde 1606, proviene del griego ἐπιδημία, *epidēmía*, de *epi* (sobre) y *demos* (pueblo), que no contiene ningún significado sobre enfermedades. Quiere decir propiamente «residencia en un lugar o país», derivada de *epidēmōō*, «yo resido en un lugar en calidad de extranjero», en oposición a *endémico*, derivada de *endēmōō*, «yo vivo en un lugar permanentemente». (2) El significado de epidemia como enfermedad que afecta a un gran número de personas en un tiempo determinado es del latín tardío, que modificó el término prestado del griego. Además de *epi* y *demos*, la etimología de epidemiología incluye *logos* (palabra, tratado) y el sufijo *ia* (cualidad).

La epidemiología tiene un cuerpo propio que se apoya en muchas otras disciplinas que investigan sobre los determinantes de la salud y la enfermedad, como son las ciencias básicas biomédicas (la biología molecular, la genética poblacional, la fisiología, la bioquímica, la patología), las ciencias clínicas médicas y quirúrgicas, la bioestadística, la demografía y la psicometría, las ciencias de la conducta y las ciencias sociales.

La definición de epidemiología que acabamos de explicar se basa en dos supuestos fundamentales. El primer supuesto es que *la aparición de la enfermedad en las poblaciones no es un fenómeno aleatorio*. La epidemiología es una ciencia empírica pragmática, en donde apenas hay leyes de la naturaleza o afirmaciones universales, pero sin embargo sí contiene una afirmación empíricamente generalizable, que «todas las enfermedades se distribuyen de modo no aleatorio» o, expresada de otro modo más concreto, «para los seres humanos, todas las enfermedades se distribuyen de modo no aleatorio mediante constructos espaciotemporales». (3)

El segundo supuesto que subyace a la epidemiología es que «las causas de los fenómenos que afectan a la salud son, en buena medida, prevenibles», pueden ser objeto de intervención para mejorar la salud.

La intervención se puede dirigir a cualquiera de los determinantes de la salud, relacionados con el huésped, el agente causal último de la enfermedad y el ambiente, la llamada tríada epidemiológica (figuras 1.a y 1.b), que interactúan entre sí aumentando o disminuyendo el riesgo de enfermedad.

1.1.1. Los fines de la epidemiología

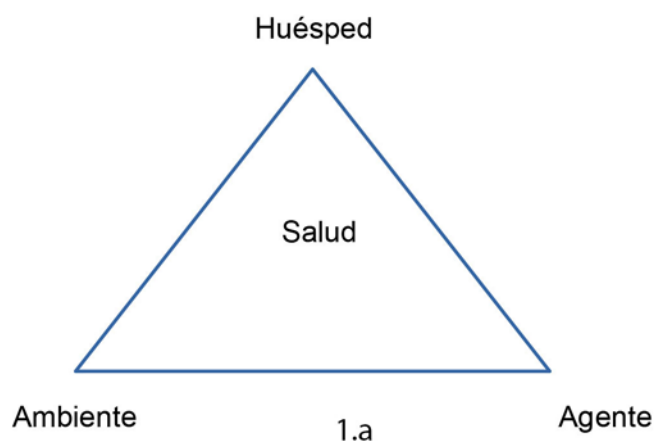
Los principales fines de la epidemiología se pueden definir desde dos perspectivas superpuestas. La primera es la perspectiva biomédica dirigida a conocer las causas de las enfermedades y su naturaleza:

- la descripción del espectro de la enfermedad,
- la descripción de su historia natural,
- la determinación de sus marcadores de riesgo, diagnósticos o de cribado,
- la identificación de las causas de las enfermedades: los factores que aumentan o reducen el riesgo de que se desarrollen,
- la predicción de las tendencias temporales y geográficas, y
- la clarificación de las formas de transmisión de las enfermedades infecciosas.

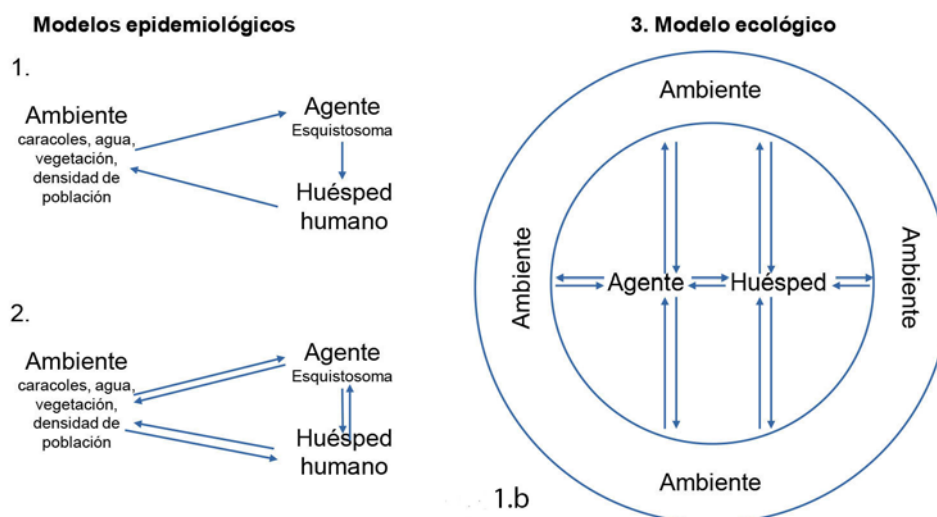
La segunda es la perspectiva de la salud pública, por la que la epidemiología, junto con otras ciencias sociales, proporciona los métodos cuantitativos y analíticos, los principios de indagación lógica y las normas probatorias básicas de:

- la medición, el diagnóstico y la proyección de las necesidades sanitarias de las poblaciones,
- la determinación de los objetivos y prioridades en la salud,
- la distribución y gestión de los recursos de la asistencia sanitaria, y
- la evaluación de las consecuencias de la enfermedad, las intervenciones sanitarias y los servicios de salud.

Figura 1. Modelos de tríadas epidemiológicas.



1.a Tríada epidemiológica clásica



1.b El modelo de tríada epidemiológica clásica, modelo 1, se ha transformado en un modelo en el que los elementos de la tríada interaccionan entre sí (modelo 2) o en otro que probablemente refleje mejor la realidad (modelo 3), en donde el agente y el huésped interaccionan continuamente dentro de un ambiente que los cubre. Las figuras de los modelos 2 y 3 toman la esquistosomiasis como ejemplo y están adaptada de una figura de Susser. (4)

1.1.2. La epidemiología y la salud pública

Según *A Dictionary of Epidemiology*,⁽⁵⁾³ la salud pública es uno de los esfuerzos organizados por la sociedad para proteger, promover y restaurar la salud de la población. Es la combinación de ciencias, habilidades y creencias orientadas al mantenimiento y la mejora de la salud de toda la población mediante acciones colectivas o sociales. Los programas, servicios e instituciones implicadas ponen el énfasis en la prevención de la enfermedad y en las necesidades de salud de la población en su conjunto. Las actividades de la salud pública cambian con los cambios en la tecnología y con los valores sociales, pero sus fines son siempre los mismos: reducir el número de enfermos, la muerte prematura, las molestias ocasionadas por la enfermedad y la discapacidad en la población. De este modo, la salud pública es tanto una institución como una disciplina y una práctica. El diccionario cita la definición de salud pública del *Informe Acheson*: la ciencia y el arte de prevenir la enfermedad, prolongar la vida y promover la salud por medio de los esfuerzos organizados de la sociedad.

Pues bien, la práctica de la salud pública se basa en el conocimiento científico de los factores que afectan a la salud y a la enfermedad, y la epidemiología constituye su ciencia básica, con el complemento de la bioestadística y de las demás ciencias ya mencionadas. La epidemiología es la rama de la salud pública que trata de descubrir las causas de la enfermedad con el fin de hacer posible su prevención. (6)

1.2. Objetivos específicos de la epidemiología

1.2.1. La identificación las *causas* de las enfermedades. El objetivo último de esta identificación es intervenir para reducir la morbilidad y la mortalidad asociadas a las enfermedades, para poner en marcha políticas preventivas. Por ejemplo, ¿cuáles son las causas de la obesidad?

1.2.2. La determinación de la *frecuencia de las enfermedades* en una población determinada. Con esto se podrá llevar a cabo una planificación adecuada de los servicios y de las instalaciones sanitarias. Por ejemplo, ¿cuál es la prevalencia de la demencia en una población determinada?

1.2.3. El estudio de la *historia natural* y del *pronóstico* de las enfermedades. De este modo, se podrá evaluar la efectividad de las intervenciones preventivas o terapéuticas aplicadas. Por ejemplo, ¿cuál es la historia natural en lo que respecta a la discapacidad y a la mortalidad de las distintas formas de la esclerosis múltiple?

³ En la sexta edición de *A Dictionary of Epidemiology*, (1) más reciente, la definición de salud pública se compone de cuatro acepciones que recogen su uso en la práctica, pero no contiene expresamente una definición conceptual. (1)

1.2.4. La evaluación de las *intervenciones preventivas y terapéuticas y de los sistemas de provisión sanitaria*. Esta evaluación permite juzgar su impacto en los resultados de salud de los pacientes implicados. Por ejemplo, ¿sirve el cribado con el antígeno prostático específico (PSA, por sus siglas en inglés) para reducir la mortalidad por cáncer de próstata?

1.2.5. El establecimiento de las bases para el *desarrollo de políticas públicas* relacionadas con problemas del medioambiente, con cuestiones genéticas y con otras consideraciones sociales y conductuales relacionadas con la prevención de la enfermedad y la promoción de la salud. Por ejemplo, ¿en qué medida es la contaminación atmosférica causada por las partículas en suspensión —PM_{2,5} y PM₁₀— perjudicial para la salud de la población?

1.3. Conceptos de la epidemiología

1.3.1. Aunque la epidemiología no investiga los mecanismos biológicos que conducen de la exposición a la enfermedad, sí puede *identificar las situaciones modificables* que afectan a la salud. La historia de muchas enfermedades infecciosas, como el cólera o la tuberculosis, muestra que la intervención sobre sus factores modificables, como la calidad del agua, el hacinamiento o el lavado de manos, puede ayudar a su control mucho antes de que se descubra su agente infeccioso.

1.3.2. Lo que distingue a la epidemiología es su *perspectiva de grupos* o poblaciones, no de individuos. La epidemiología es una disciplina que se basa en la comparación de grupos, por lo que la inclusión de un grupo de control es crucial para establecer inferencias causales (ver los capítulos de *Estudios de intervención e Inferencia causal*).

1.3.3. Todos los estudios epidemiológicos, sin excepción, están expuestos a *problemas de sesgos y confusión* que pueden comprometer su validez interna, especialmente los estudios no experimentales, que son los más comunes.

La mayoría de los estudios epidemiológicos se efectúan en *poblaciones mixtas y diversas*. Esto puede amenazar su validez interna, al estar siempre presente el riesgo de confusión, que los grupos en estudio no sean completamente comparables en las variables que influyen en la exposición y en el desenlace estudiados.

1.3.4. Casi nunca es factible trabajar con muestras representativas de la población general de interés. Lo más importante es la validez interna de los estudios y que las muestras estén tomadas de una población bien definida. Pero en todos los estudios hay que tener en cuenta la validez externa de sus resultados, hasta qué punto son generalizables de la población fuente a otras poblaciones distintas.

1.3.5. Casi nunca basta un único estudio para obtener suficientes pruebas que justifiquen una intervención. Los hallazgos relevantes tienen que reproducirse y confirmarse por otros estudios posteriores. Los estudios con resultados negativos (en donde no hay pruebas del efecto) tienen tanto valor como los de resultados positivos.

1.4. Epidemiología y prevención

La prevención se define como las acciones dirigidas a la erradicación, eliminación o minimización del impacto de la enfermedad y la discapacidad o, si ninguno de estos fines fuera posible, al retraso en el progreso de la enfermedad y la discapacidad. (1)

La epidemiología permite identificar los grupos de población que tienen mayor riesgo de contraer determinadas enfermedades. Esto permite centrar en ellos los esfuerzos preventivos y terapéuticos. Además, se puede estudiar en esos grupos los factores específicos que aumentan su riesgo para poder intervenir en ellos. Esos factores pueden ser no modificables (el sexo, la edad, la etnia) o modificables (obesidad, diabetes, contaminación, estilos de vida...).

Hay tres tipos principales de prevención:

- la prevención primaria, que previene el desarrollo inicial de la enfermedad, por medio de medidas como la vacunación, el suministro de preservativos, la quimioprofilaxis o la práctica del ejercicio físico;
- la prevención secundaria, que busca la detección precoz de una enfermedad que todavía no se ha expresado clínicamente con objeto de reducir su gravedad y sus complicaciones, como, por ejemplo, el cribado del cáncer colorrectal o de la hipertensión arterial, o que trata de reducir la incidencia de sucesos cardiovasculares en pacientes que ya han sufrido un primer suceso mediante medicamentos y medidas para promover estilos de vida saludables, y
- la prevención terciaria, que busca reducir el impacto de la enfermedad, como, por ejemplo, la rehabilitación tras un ictus.

A estos tipos podemos añadir otros dos tipos de prevención:

- la prevención primordial, que busca prevenir la aparición de los determinantes poblacionales de las enfermedades y así inhibir el establecimiento de los factores ambientales, económicos, sociales, conductuales y culturales que aumentan su riesgo. Actúa a nivel sistémico, no a nivel personal, que sería el campo de aplicación de la prevención primaria. Dentro de la prevención primordial estarían medidas como la potabilización del agua y el

saneamiento de las aguas residuales, las leyes antitabaco, la promoción de la salud en la infancia desde la fase prenatal o el desarrollo de ciudades más habitables y limpias.

- la prevención cuaternaria, la identificación de individuos y grupos con riesgo de sobrediagnóstico o sobremedicación para la prevención de la iatrogenia y para mejorar la seguridad del paciente. Este tipo de prevención se expresa en la reducción de intervenciones médicas excesivas que son fruto de la construcción social de la enfermedad y que están muy influidas por los intereses de compañías con ánimo de lucro y por la llamada mercantilización de la enfermedad.

Hay dos enfoques posibles de la prevención, (7)

- el enfoque poblacional, de salud pública, es decir, la prevención del riesgo en la población en su conjunto, como el establecimiento de un límite para la cantidad de sal del pan, las leyes que prohíben el consumo de tabaco en interiores, o las medidas para reducir el consumo de ácidos grasos *trans*, y
- el enfoque individual, clínico (o de alto riesgo), que busca la prevención del riesgo en sujetos con enfermedad o factores de riesgo manifiestos. Por ejemplo, el tratamiento correcto de la hipertensión arterial, un factor de riesgo importante de sucesos cardiovasculares, permite reducir su incidencia.

En la práctica, el enfoque poblacional puede ser más eficaz si una buena proporción de los casos de enfermedad se produce en personas sin niveles altos de sus factores de riesgo, como sucede en la enfermedad cardiovascular. Eso sí, las medidas de prevención poblacionales tienen que ser baratas, no invasivas e indoloras. En general, las medidas poblacionales tienen que combinarse con la atención sanitaria individual para quienes ya tienen factores de riesgo establecidos.

1.5. La epidemiología y la práctica clínica

La epidemiología no solo sienta las bases de la salud pública, también es la base científica de la medicina basada en pruebas (mal llamada medicina basada en la evidencia). La información científica que se dispone de la descripción, la etiología, el diagnóstico, el pronóstico y el tratamiento de las enfermedades proviene casi siempre de estudios efectuados en grupos de pacientes.

En el ámbito clínico, el conocimiento que aporta la epidemiología para el control de los problemas de salud no se aplica a las poblaciones, como en la salud pública, sino a los pacientes individuales. Por ejemplo, un médico que tenga que tratar a un paciente con un infarto agudo de miocardio conoce que determinados tratamientos (angioplastia, betabloqueantes, estatinas) pueden mejorar su pronós-

tico porque esos tratamientos han demostrado su beneficio en ensayos controlados y aleatorizados de pacientes similares.

La epidemiología clínica es la aplicación de los conceptos y métodos epidemiológicos al estudio de diversos aspectos en poblaciones de pacientes.

1.6. El enfoque epidemiológico

El razonamiento epidemiológico se compone de tres fases:

- la primera fase consiste en determinar si hay asociación entre la exposición a un factor o una característica con la presencia de la enfermedad que se está estudiando, mediante el estudio de las características de los grupos y de los individuos;
- después, una vez encontrada la asociación, surge la pregunta de si esa asociación es verdaderamente causal o no, es decir, si se puede establecer una inferencia causal entre la exposición y la enfermedad, y, por último,
- intervenir sobre la causa para eliminar o reducir la enfermedad.

El enfoque epidemiológico se ha aplicado, por ejemplo, en el estudio de la relación entre el tabaquismo pasivo y la mortalidad perinatal, de niños menores de un año, y la mortalidad coronaria y por cáncer de pulmón en adultos. La comprobación de la asociación dio lugar a una inferencia causal clara, lo que sirvió de base para que las autoridades de salud pública tomaran medidas para prohibir fumar en espacios cerrados.

2. Historia de la epidemiología

El conocimiento de la historia nos permite saber de dónde venimos y cómo fue que llegamos hasta aquí. También nos puede ayudar a saber a dónde vamos.

Las primeras ideas en la historia plasmadas por escrito que intentan buscar una explicación de las causas de las enfermedades las encontramos en Hipócrates (*circa* 460-370 a. C.), en particular en sus libros *Epidemia I*, *Epidemia III* y *Aires, aguas y lugares*. Se le considera el padre de la medicina y de la epidemiología.

Hipócrates puso el énfasis en la influencia del medio ambiente (a través del aire, el agua y los lugares) —un enfoque ecológico que cobra mucha importancia hoy en día— y de la conducta individual en la salud. Por ejemplo, señaló que el

paludismo y la fiebre amarilla eran más comunes en zonas pantanosas. Por otra parte, su enfoque fue racionalista, es decir, era un enfoque contrario a la influencia de la ira de los dioses, a la superstición en la salud, es decir, a causas sobrenaturales.

Sin embargo, el enfoque hipocrático buscaba las causas específicas de cada enfermo, tomado individualmente, con el supuesto de que cada caso era distinto, mientras que apenas tenía interés en los patrones poblacionales de enfermedad. Hipócrates creía que las enfermedades eran producidas por algún tipo de contaminación ambiental, los miasmas, el conjunto de emanaciones fétidas de suelos y aguas impuras. Pero la teoría miasmática de la enfermedad, para explicar por qué no enfermaba todo el mundo, requería que los miasmas solo enfermaran a aquellos predispuestos, a quienes tenían exceso o déficit de sus humores o fluidos corporales⁴. Lo que daba lugar a que cada caso fuera tratado de forma distinta.

El enfoque racionalista hipocrático desapareció de Occidente con el colapso de la civilización clásica y el oscurantismo que vino después. Entre los siglos III y XV la salud y la enfermedad se concibieron como castigo o perdón divinos. (8) El legado teórico de Hipócrates se mantuvo, sin embargo, en el Oriente Próximo, en donde se tradujeron al árabe sus textos. Ibn an-Nafis (*circa* 1210-1288), médico de Damasco que se estableció en El Cairo y descubrió la circulación menor casi 300 años antes que Miguel Servet, fue un seguidor de Hipócrates y se le puede considerar un epidemiólogo de su tiempo. (9)

¿Cuándo aparece el término epidemiología? En 1598 Quinto Tiberio Angele-rio, un médico italiano, publicó en Madrid un libro con el título de *Epidemiologia sive tractatus de peste*, escrito en latín, pero que contiene un apéndice en español con las medidas para controlar una epidemia. En este apéndice aparece por primera vez la palabra epidemiología. (10) Doscientos años después, en 1802, Joaquín de Villalba publica en español el libro *Epidemiología española. O historia cronológica de las pestes, los contagios, epidemias y epizootias que han acaecido en España desde la venida de los cartaginenses hasta el año 1801*, que describe el estudio de todas las epidemias históricas hasta entonces, no solo la de la peste. (11)

El desarrollo histórico de la epidemiología va estrechamente ligado al estudio de las epidemias que han azotado la humanidad en tiempos modernos. (12) Alfredo Morabia establece cuatro periodos sucesivos en la historia de la disciplina: la epidemiología incipiente, la temprana, la clásica y la moderna. (13,14)

⁴ Los humores eran la bilis negra, la bilis amarilla, la flema y la sangre. Tenían que ver con los cuatro elementos, el fuego, el aire, el agua y la tierra, y se relacionaban también con las cualidades de calor, frío, humedad y sequedad. Esta visión de la enfermedad se mantuvo vigente en Occidente hasta el siglo XIX.

2.1. Epidemiología incipiente

La epidemiología incipiente se refiere al periodo en que las investigaciones de las causas de las enfermedades ya incorporan el pensamiento poblacional y las comparaciones de grupos, aunque sin ninguna teoría subyacente que las sustente.

A continuación, y sin ánimo de exhaustividad, nos vamos a detener en nombrar varias personalidades que fueron relevantes en el desarrollo de esta fase incipiente de la epidemiología. Con dos puntualizaciones previas. Las personas que se nombran no fueron los únicos protagonistas, pues las personas solo reflejan las tendencias de su época, aunque el relato de lo ocurrido se personalice en ellos; en segundo lugar, las aportaciones de esas personas no fueron lineales y progresivas, porque en una disciplina como la epidemiología, como en todas las disciplinas científicas, las ideas imperantes reflejan las corrientes dominantes del pensamiento de su época, corrientes cambiantes con las sociedades y en el tiempo (tabla 1).

Tabla 1. Recorrido cronológico de la historia de la epidemiología

Periodo	Protagonistas	Perspectivas
Antes del pensamiento poblacional		
circa 460-370 a. C.	Hipócrates	Influencia del medioambiente y de la conducta individual sobre la salud
Epidemiología incipiente (siglo XV-finales siglo XIX)		
1546	Girolamo Fracastoro	Concepto de enfermedad contagiosa
1580	Guillaume de Baillou	Descripción de enfermedades contagiosas
1642	Wu Youke	Teoría de la existencia de agentes patógenos transmisibles
1662	John Graunt	Tablas de vida
1676	Thomas Sydenham	Descripción nosológica de muchas enfermedades
1700	Bernardino Ramazzini	Descripción de enfermedades laborales
1747	James Lind	Primer ensayo clínico, para el tratamiento del escorbuto
1796	Edward Jenner	Vacuna contra la viruela
1835	Pierre-Charles-Alexandre Louis	Comparación entre grupos semejantes y pensamiento poblacional

Tabla 1. Recorrido cronológico de la historia de la epidemiología (continuación)

Periodo	Protagonistas	Perspectivas
1853	John Snow	Muertes por cólera y fuentes de suministro de agua
1861	Ignaz Semmelweis	Higiene de manos y fiebre puerperal
Década de 1880	Lous Pasteur y Robert Koch	Teoría microbiológica de la enfermedad
Siglo XIX	Louis René Villermé, Frederich Engels, Rudolph Virchow y Émile Durkheim	Inicios de la epidemiología social
Epidemiología temprana (finales siglo XIX-1945)		
1838	William Farr	Conceptos de tasa, tiempo-persona, efecto dosis-respuesta, inmunidad de grupo y otros
1939	Austin Bradford Hill	Distinción entre riesgo y tasa
Década de 1910	Janet Lane-Claypon y Wilhelm Weinberg	Primeros estudios de cohortes de la edad moderna
1918	William Henry Welch y Wade Hampton Frost	Fundación de la Johns Hopkins University School of Hygiene and Public Health y dirección del departamento de Epidemiología, respectivamente
1920	Joseph Goldberger y Edgar Sydenstricker	Carencias vitamínicas como causa de enfermedad
1926	Janet Lane-Claypon	Primer estudio de casos y controles
1928-1935	Major Greenwood	Primer catedrático de Epidemiología en el Reino Unido que publicaría en 1935 el que se considera el primer libro de texto de epidemiología científica <i>Epidemics and crowd-diseases</i>
1941	Joseph Asbury Bell	Publicó un ensayo con ocultación del reparto aleatorio
Finales del siglo XIX-primera mitad siglo XX	Florence Nightingale, W. E. B. Du Bois, Alice Hamilton y Salvador Allende	Contribuciones a la epidemiología social
Epidemiología clásica (1945-finales de la década de 1970)		
1946	Joseph Berkson	Desarrollo del concepto de sesgo de selección
1948	Philip D'Arcy Hart, Austin Bradford Hill y otros miembros del Streptomycin in Tuberculosis Trials Committee	Publicación del que se considera el primer ensayo controlado y aleatorizado de la época actual, el estudio de la estreptomicina como tratamiento de la tuberculosis pulmonar

Tabla 1. Recorrido cronológico de la historia de la epidemiología (continuación)

Periodo	Protagonistas	Perspectivas
Década de 1950	Richard Doll y Austin Bradford Hill	Publicaron un estudio de casos y controles sobre el tabaco y el cáncer de pulmón e iniciaron el <i>British Cohort Study</i> , un estudio de cohortes en médicos británicos sobre los efectos del tabaco
1951	Jerome Cornfield	La <i>odds ratio</i> puede estimar la razón de riesgos si se cumple el supuesto de enfermedad rara
1952	E. Cuyler Hammond y Daniel Horn	Se inició el <i>Hammond and Horn Study</i> , con casi 200 000 participantes, por la American Cancer Society, para evaluar los efectos del tabaco
Década de 1950	Edward H. Simpson, Nathan Mantel y William Haenszel	Avances en el desarrollo del concepto y análisis de la confusión
Décadas de 1950 y 1960	Ernst Wynder, Abraham Lilienfeld, Saxon Graham y David J. Newell	Polémica sobre el sesgo de información
Desde 1956 a 1976	Morris, MacMahon <i>et al.</i> , McMahon y Pugh, Susser y Lilienfeld	Libros de texto de epidemiología
1964	Surgeon General de EE. UU.	Publicación del informe <i>Smoking and Health</i> , sobre la asociación entre fumar y el cáncer de pulmón
1965	Austin Bradford Hill	Publica una serie de características para la inferencia causal
1976-1977	Richard Peto <i>et al.</i>	Publican dos artículos sobre los métodos y el análisis estadístico de los ensayos controlados y aleatorizados
1979	E. Cuyler Hammond	Desarrollo del concepto de interacción en la relación entre el tabaco, el amianto y el cáncer de pulmón
1979	David L. Sackett	Intento de catalogar los sesgos en la investigación epidemiológica
Epidemiología moderna (finales de la década de 1970 hasta la actualidad)		
1974-1976	Kenneth J. Rothman	Desarrolla el concepto de interacción y lo enmarca dentro de las causas de las enfermedades

Tabla 1. Recorrido cronológico de la historia de la epidemiología (continuación)

Periodo	Protagonistas	Perspectivas
1976	Olli S. Miettinen	Establece el vínculo entre riesgo y tasa y entre los estudios de casos y controles y los de cohortes; cuestiona el supuesto de enfermedad rara como condición para que la <i>odds ratio</i> estime la razón de riesgos o la razón de tasas
1977	Karen T. Copeland <i>et al.</i>	Profundización de los conceptos de sesgo de información y de errores diferenciales y no diferenciales
1986	Sander Greenland y James M. Robins	Profundización del concepto de confusión, tomado como que las poblaciones que se comparan sean intercambiables
Década de 1980- década de 2000	Sackett, Feinstein, Fletcher, Fletcher, Wagner, Weiss, Guyatt	Epidemiología clínica, medicina basada en la evidencia, GRADE (<i>Grading of Recommendations Assessment, Development and Evaluation</i>)
Década de 1980-actualidad	Richard Peto <i>et al.</i>	Revisiones sistemáticas y metanálisis en enfermedades cardiovasculares y en cáncer
1982-2021	Kleinbaum, Rothman, Greenland, Lash, Gordis, Szklo, Nieto, Celentano	Múltiples libros de texto de epidemiología
1992-1996	Mervyn Susser, Ezra Susser, Geoffrey Rose, Nancy Krieger, Sally Zierler y Neil Pearce	Revitalización de la epidemiología social
1993	Iain Chalmers y Peter C. Gøtzsche	Fundación de la Colaboración Cochrane
1999	Anthony J. McMichael	Propuestas de renovación de la epidemiología
2004	Miguel A. Hernán <i>et al.</i>	Profundización del concepto de sesgo de selección

- Girolamo Fracastoro (1478-1553), médico italiano de Verona, publicó en Venecia en 1546 el libro *De contagione et contagiosis morbis et eorum curatione*, en el que se estableció por primera vez el concepto de enfermedad contagiosa, describió las enfermedades contagiosas de entonces (peste, lepra, tisis, sarna, rabia, erisipela, viruela, ántrax, tracoma, tifus exantemático y sífilis) y clasificó los contagios en tres tipos, directo, mediante fómites y a distancia. También propuso la higiene personal y ambiental para pre-

venir los contagios. Hizo otras dos contribuciones importantes: separó el concepto de infección, como causa, del de epidemia, como consecuencia; y se cree que fue el primero en establecer que enfermedades específicas son la consecuencia de contagios específicos. Uno de sus contemporáneos, Girolamo Cardano, llegó a decir en 1557 que «las semillas de la enfermedad eran animales diminutos capaces de reproducirse», con lo que inició la teoría del contagio, que se desarrollaría en los siglos posteriores. (15)

- Guillaume de Baillou (1538-1616), médico francés que revitalizó la medicina de Hipócrates en el Renacimiento, publicó en 1580 el libro *Epidemiorum, Sobre las epidemias*, en el que hacía una descripción completa de las epidemias de sarampión, difteria y peste bubónica que aparecieron en Europa entre 1570 y 1579. Fue el primero en describir la tos ferina y en definir el reumatismo en su sentido actual.
- Poco después, en el Lejano Oriente, al final de la Dinastía Ming, Wu Youke (1582-1652) postuló la teoría de que algunas enfermedades eran producidas por agentes transmisibles, lo que llamó factores pestilentes, lo que plasmó en su libro *Wen Yi Lun*, Tratado de la pestilencia (de las enfermedades epidémicas), publicado en 1642, tras observar las epidemias que azotaron a la población de su alrededor. Wu Youke inició una escuela china que desarrolló sus teorías, pero estuvo desconectada de Occidente casi por completo hasta el siglo XIX. (16)
- John Graunt (1620-1674) publicó en 1662 *Natural and political observations made upon the bills of mortality*. Analizó los boletines de mortalidad publicados en Londres desde 1603 y a partir de ellos construyó el embrión de lo que después serían las tablas de vida (tablas de mortalidad), que permitieron calcular la probabilidad de supervivencia en los años en que hubo una epidemia de peste.⁵ Encontró que el número de nacimientos era mayor en varones que en mujeres, en una razón de 14 a 13, que uno de cada tres niños morían antes de los 5 años, y que había variaciones en la probabilidad de supervivencia según el sexo, el lugar de residencia, la estación meteorológica y la edad. (17) Fue precursor de la epidemiología y se le considera el fundador de la bioestadística y el primer demógrafo.

⁵ No fueron John Graunt ni William Petty —economista y amigo suyo que publicó trabajos sobre los patrones demográficos de la población inglesa— los que establecieron las tablas de vida. En realidad, quienes lo hicieron fueron las compañías aseguradoras para fijar los precios de los seguros de vida, muy comunes en Francia desde el siglo XVI y en Inglaterra y Gales desde el siglo XVII, a través de las asociaciones de socorros mutuos. (8)

- Thomas Sydenham (1624-1689) publicó en 1676 *Observationes medicae, Observaciones médicas sobre la historia y curación de las enfermedades agudas*, una descripción nosológica de muchas enfermedades, efectuada a partir de su observación, análisis y comparación, que sentó las bases del estudio de la historia natural de la enfermedad. Defendía que las enfermedades se pueden clasificar al igual que los botánicos clasifican las plantas. Se le considera el padre de la medicina inglesa o el Hipócrates inglés, aunque de forma póstuma, ya que en vida no tuvo ese reconocimiento por sus contemporáneos. Intervino en una época de transición de concebir la enfermedad, en general, como un ente único, a clasificarla en distintas entidades específicas.

Fue un defensor de la teoría miasmática de la enfermedad, por lo que apoyó el valor terapéutico del aire limpio. También defendió el ejercicio y una buena dieta, en contra de muchos colegas de su tiempo.

- Bernardino Ramazzini (1633-1714) fue un médico italiano que publicó en 1700 *De morbis artificum diatriba, Tratado sobre las enfermedades laborales*, una descripción minuciosa de las enfermedades de 54 ocupaciones distintas. Se le considera el precursor de la medicina y de la epidemiología laboral y también de la higiene industrial.
- James Lind (1716-1794), un médico escocés, cuando estaba en la marina, en 1747, practicó un ensayo clínico (el primero conocido de la historia) en 12 marineros del barco Salisbury enfermos de escorbuto, divididos en 6 grupos de 2. El grupo que recibió 2 naranjas y un limón por día fue el único que se curó. Lo publicó en *A treatise of the scurvy, in three parts. Containing an inquiry into the nature, causes, and cure, of that disease. Together with a critical and chronological view of what has been published on the subject*, en 1753. Hizo muchas contribuciones epidemiológicas: estudiar las enfermedades en las poblaciones, formular preguntas epidemiológicas clásicas, efectuar observaciones de los cambios ocurridos en el tiempo y aplicar un diseño experimental.
- Edward Jenner (1749-1823), un médico rural inglés, conocía por observaciones provenientes de Inglaterra, Hungría, Turquía, China y América que las personas infectadas de la viruela de las vacas no se contagiaban de la viruela y él mismo lo comprobó en su experiencia clínica. En mayo de 1796 vio que una mujer lechera tenía en las manos lesiones activas de la viruela de las vacas. Inoculó el material de las lesiones a un niño de 8 años, que tuvo un cuadro leve. Cuando se recuperó le inoculó material de lesiones

de viruela humana y el niño no enfermó.⁶ Fue el inicio de la vacunación contra la viruela, que gracias a ello se erradicó en 1978.

- Pierre-Charles-Alexandre Louis (1787-1872) fue un médico francés que demostró que las sangrías, muy populares entonces como tratamiento generalizado de cualquier fiebre, eran inútiles en el tratamiento de la neumonía y podían ser perjudiciales. Esto lo publicó en 1835 en el libro *Recherches sur les effets de la saignée dans quelques maladies inflammatoires et sur l'action de l'émétique et des vésicatoires dans la pneumonie*. Louis defendió el uso del *método numérico*, lo que hoy llamaríamos la estadística, para conocer las leyes de la mortalidad. Y para ello defendió ante la resistencia de sus colegas dos conceptos básicos de la epidemiología: la comparación entre grupos lo más semejantes posible excepto en el tratamiento y el pensamiento poblacional. (20) Por esos motivos, se le puede considerar un pionero de la epidemiología clínica y un precursor de la medicina basada en pruebas.
- John Snow (1813-1858), un médico inglés, aplicó criterios de epidemiología descriptiva y analítica para estudiar una epidemia de cólera en Londres en 1853, por lo que se le conoce como uno de los que más ha contribuido a la epidemiología. Creía firmemente en que pequeños animálculos transmitían el cólera por contagio. Estudió las tasas de muerte por cólera en Londres según las viviendas de los fallecidos recibieran agua de la compañía Lambeth o de Southwark and Vauxhall. Como no disponía de datos sobre la población, utilizó como denominadores las viviendas servidas por cada compañía. Encontró que la mortalidad en las casas servidas por Southwark and Vauxhall Water Company —aguas abajo del Támesis, una vez que habían recibido el vertido de las aguas fecales de Londres— fue 8,5 veces superior a la de las casas servidas por Lambeth Water Company —aguas arriba, que filtraba las aguas—. El cólera se transmitía por el agua contaminada, no por el aire.
- A Ignaz Semmelweis (1818-1865), un cirujano y obstetra del imperio austrohúngaro, se le conoce como el *salvador de las madres*, por haber descubierto que la higiene de manos reducía las fiebres puerperales, lo que recogió en su obra de 1861 *De la etiología, el concepto y la profilaxis de la fiebre puerperal*.

⁶ Esta inoculación intencionada de las lesiones de viruela humana a un niño es cuestionable desde el punto de vista ético. Eso no quita que haya circunstancias en que la inoculación del agente infeccioso pueda estar justificada, como se discutió para los ensayos de vacunas contra la COVID-19. (18,19)

Semmelweis, cuando estuvo a cargo de la Primera Clínica Obstétrica del Hospital General de Viena, observó que la mortalidad de su clínica era del 16 %, mientras que la de otra clínica obstétrica del mismo hospital, la Segunda Clínica Obstétrica, tenía una mortalidad del 7 %. Se daba la circunstancia de que en su clínica los médicos y los estudiantes, antes de explorar a sus pacientes, pasaban por la sala de autopsias, en donde estudiaban a mujeres fallecidas por fiebre puerperal, mientras que la segunda clínica era atendida por matronas, sin ninguna relación con las autopsias. Postuló que los médicos y los estudiantes, cuando exploraban a las mujeres, al producirles pequeños traumas les transmitían partículas que ocasionaban la enfermedad. Se reafirmó en esta hipótesis cuando vio a un amigo y colega suyo morir después de una infección cutánea tras pincharse accidentalmente al hacer una autopsia y con hallazgos necróticos semejantes a los de las mujeres que morían por fiebre puerperal. Y obligó a que todos se lavaran las manos y se cepillaran las uñas después de haber estado en la sala de autopsias, antes de explorar a las mujeres parturientas. Como consecuencia, la mortalidad en su clínica descendió rápidamente.

Años después, se comprobó que las fiebres puerperales se producían predominantemente por infección estreptocócica, al igual que el virus de la viruela se descubrió después del experimento de Edward Jenner y el *Vibrio cholerae* después de los trabajos de John Snow. Estos ejemplos muestran que es posible aplicar políticas preventivas sin que se conozca la causa biológica de la enfermedad.

- Louis Pasteur (1822-1895) fue un químico y microbiólogo francés que acabó con la teoría de la generación espontánea de las enfermedades infecciosas y desarrolló, en su lugar, la teoría microbiana. Su mayor contribución a la epidemiología fue la experimentación para demostrar la eficacia de una vacuna atenuada contra el carbunco en ovejas. Este experimento se compuso de tres grupos: uno experimental vacunado con bacterias atenuadas, en el que todos sobrevivieron tras inocularles el carbunco; un primer grupo de control no vacunado, que murieron tras inocularles el carbunco, y un segundo grupo de control sin vacunación ni inoculación, en el que las ovejas se mantuvieron como estaban). En este experimento también demostró que las bacterias —las esporas de *Bacillus anthracis* ya se habían identificado— eran la causa de la enfermedad.⁷

⁷ Louis Pasteur hizo otras contribuciones importantes. Creó una vacuna contra el cólera aviar. Y estaba diseñando una vacuna atenuada contra la rabia a partir de la desecación del tejido nervioso de conejos infectados, que solo había probado en perros, cuando en 1885 un niño, Joseph Meister, fue mordido 14 veces por un perro rabioso, por lo que estaba condenado a morir cuando desarrollara la enfermedad. Pasteur no era mé-

- Robert Koch (1843-1910) fue un médico y microbiólogo alemán que aisló los gérmenes causantes del carbunco, la tuberculosis y el cólera. Descubrió que la filtración del agua reducía la transmisión del cólera y también contribuyó al entierro de la teoría de la generación espontánea. Formuló los famosos postulados de Koch (o de Henle-Koch) basándose en conceptos previos definidos por Jakob Henle con objeto de establecer una relación causal entre un microbio y una enfermedad, base de la teoría microbiana de la enfermedad.

Un episodio curioso ilustra las disputas de la época. El higienista alemán Max von Pettenkofer (1818-1901), una autoridad muy reputada en la salud pública de entonces, quiso reconciliar la teoría miasmática de la enfermedad con la teoría del contagio en el origen del cólera. Sostenía que para causar la enfermedad el germen causante del cólera tenía que transformarse mediante su migración por suelos porosos, lo que podría ser un anticipo del concepto futuro de interacción. Rechazaba ardientemente la teoría monocausal del germen, que defendía Koch. Para refutar esa teoría llegó a ingerir directamente una solución con el germen. Von Pettenkofer no pudo demostrar su teoría porque no recurrió a comparaciones de grupos. (13)

Este nuevo paradigma, en el que la causa de cada enfermedad contagiosa podía explicarse por un único agente infeccioso, se amplió a otras enfermedades no infecciosas cuyos agentes causales específicos podían ser carcinógenos, deficiencias de vitaminas o alteraciones genéticas. La teoría microbiana de la enfermedad sepultó para siempre la teoría miasmática de la enfermedad que, a pesar de basarse en escasas pruebas científicas, había prevalecido hasta entonces, pero que también sirvió para poner en marcha medidas de higiene de las aguas y del aire que fueron a menudo beneficiosas para mejorar la salud de las poblaciones.⁸

Los trabajos de Lind, Louis, Semmelweis y Snow fueron acogidos con escepticismo en su tiempo. Se cuestionaba que los grupos como tales fueran entidades dignas de estudio y, en caso afirmativo, si los grupos eran comparables o no, anticipando lo que hoy denominamos confusión. En aquellos tiempos, faltaba una teoría de comparación de grupos. De hecho, Snow cayó casi en el olvido hasta que fue recuperado para la historia de la epidemiología en la década de 1930 cuando su aportación encajaba en el paradigma bacteriológico dominante. (21)

co y podía tener problemas legales por inocular su vacuna al niño, pero decidió hacerlo. El niño no contrajo la enfermedad, la vacuna le salvó la vida. El descubrimiento de la vacuna antirrábica es uno de los hitos más importantes de la historia de la medicina.

⁸ Aunque también pudieron ser perjudiciales. Se cuenta que en Londres limpiaban la suciedad de las calles, que incluía las heces humanas, y las arrojaban al Támesis, que era la fuente del agua para beber. En tiempos del cólera, eso podía perpetuar la epidemia. (21)

Pero en esta fase de epidemiología incipiente también se encuentran autores de trabajos que constituyen el embrión de lo que más adelante será la epidemiología social: Louis René Villermé, *De la mortalité dans les divers quartiers de la ville de Paris* (1830), Frederick Engels, *The condition of the working class in England* (1845), Rudolph Virchow, *Informe sobre la epidemia de tifus en Alta Silesia* (1848) y Émile Durkheim, *Le suicide* (1897). En estos trabajos se ponía el énfasis en la relación entre las enfermedades y las características de la sociedad: las condiciones de vida de las ciudades, la pobreza o las ocupaciones peligrosas.

2.2. Epidemiología temprana

En esta fase ya se establecen algunos elementos de la teoría epidemiológica, la epidemiología se convierte en una disciplina académica y se empiezan a publicar libros de texto. Abarca desde finales del siglo XIX hasta el final de la Segunda Guerra Mundial, en 1945.

Por ejemplo, William Farr (1807-1883), médico epidemiólogo inglés, contemporáneo y colaborador de John Snow, considerado uno de los fundadores de la bioestadística, se propuso comprender por qué el cólera, que tenía una letalidad de casi el 50 %, infundía más terror en la gente que la tuberculosis, que tenía una letalidad de casi el 100 %. La mediana de supervivencia después del diagnóstico del cólera era de 7 días, mientras que la de la tuberculosis era de 2 años. El riesgo de muerte por cólera era del 47 % en 7 días y el de morir por tuberculosis del 100 % en un periodo superior a 2 años. Hacía falta una medida que expresara la velocidad con que la gente se moría cuando contraía la enfermedad, es decir, que expresara el riesgo por unidad de tiempo. En el caso del cólera, 47 de cada 100 morirían en una semana, mientras que, en el caso de la tuberculosis, 100 % en 2 años supone, aproximadamente, 1,0 por 100 en una semana. La medida que buscaba era la tasa, la medida por unidad de tiempo. La velocidad a la que moría la gente por tuberculosis era 47 veces inferior a la del cólera, lo que reflejaba la percepción de la población.

Farr hizo otras muchas contribuciones epidemiológicas, como el desarrollo de conceptos como el de tiempo-persona, la relación entre la tasa de muerte y la probabilidad de morir, el efecto dosis-respuesta, la necesidad de grupos muy numerosos para poder hacer inferencias causales, la inmunidad de grupo, la relación entre la prevalencia y la incidencia y la duración de la enfermedad y la direccionalidad de los estudios. (15)

Los conceptos de riesgo y de tasa se refinaron poco después. En Inglaterra hubo una polémica sobre cómo juzgar la campaña de vacunación contra la fiebre tifoidea en los soldados del ejército. A los soldados se les inculcaba la vacuna

en distintos momentos después del comienzo de la campaña. Por una parte, una comisión de militares quería definir a todos los soldados como inoculados desde el principio, siempre que estuvieran inoculados al final de la campaña. Pero esto tenía el inconveniente de que el tiempo sin enfermedad que transcurría desde el inicio de la campaña hasta la inoculación se podía atribuir falsamente al efecto de la vacuna, con lo que se exageraba su efecto. Por otra parte, un epidemiólogo, Major Greenwood (1880-1947), y un estadístico, Udny Yule (1871-1951), querían considerar solo como inoculados a los que lo estuvieran al inicio de la campaña, de modo que los que se inoculaban después figuraban como no vacunados. De este modo, se subestimaba el efecto de la vacunación. Entonces Austin Bradford Hill (1897-1991) propuso la medida de tiempo-persona para evitar las limitaciones de los dos enfoques y asignar fielmente los tiempos de seguimiento a la inoculación o a su ausencia.

Wilhelm Weinberg (1862-1937) hizo el primer estudio de cohortes sistemático para evaluar el riesgo de muerte hasta los 20 años de niños hijos de padres fallecidos por tuberculosis, en comparación con hijos de padres fallecidos por otras causas. Recurrió a un estudio retrospectivo en donde vinculó los certificados de defunción con los datos del registro civil de su estado, Baden-Württemberg, en Alemania. Generó dos cohortes, la de 18 052 hijos de personas que habían muerto de tuberculosis entre 1873 y 1889, y la de 7574 hijos de una muestra de personas que habían fallecido por otras causas durante el mismo periodo. Encontró que quienes fallecieron por tuberculosis tuvieron menos hijos que los demás y que sus hijos murieron a edades más tempranas. Lo publicó en 1913.

Poco antes, Janet Lane-Claypon (1877-1967) había publicado en Berlín otro estudio de cohortes retrospectivo más pequeño en el que tomó los datos de un registro de una clínica de planificación familiar para mostrar que los lactantes alimentados con leche materna crecían más rápido que los alimentados con leche envasada.

En 1918, William Henry Welch fundó la Johns Hopkins University School of Hygiene and Public Health y encargó a Wade Hampton Frost la dirección del departamento de Epidemiología. Este hecho se considera el comienzo de la epidemiología y la salud pública académica. En 1924 se creó en España la Escuela Nacional de Sanidad para formar especialistas en sanidad pública, entre cuyas disciplinas se señala a la epidemiología.

En 1920, Joseph Goldberger (1874-1929) y Edgar Sydenstricker (1881-1936), de Carolina del Sur, publicaron un estudio sobre los efectos de la dieta y de los ingresos económicos en la etiología de la pelagra. Para ello entrevistaron a trabajadores de las fábricas de algodón y a sus familias y les siguieron durante el verano para vigilar la aparición de pelagra. Utilizaron la restricción de la muestra como método

para mejorar la comparabilidad de los grupos. Encontraron en la monotonía de la dieta la causa de la enfermedad, pues los ingresos insuficientes y la falta de acceso a alimentos frescos obligaban a las familias a subsistir básicamente con el maíz. En el año 1929 estaba previsto conceder a Joseph Goldberger el premio Nobel por su contribución a la investigación sobre las vitaminas, pero falleció ese mismo año.

En 1926, la ya mencionada Janet Lane-Claypon publicó un estudio que hoy definiríamos como de casos y controles, en donde entrevistó a 500 mujeres hospitalizadas por cáncer de mama y a 500 mujeres hospitalizadas por enfermedades distintas al cáncer que constituyeron el grupo control. Todas fueron entrevistadas del mismo modo. Observó que las mujeres con cáncer de mama tenían en promedio menos hijos que las mujeres del grupo control (3,5 frente a 5,5, respectivamente). En este estudio, la autora llamó la atención al hecho de que las mujeres con cáncer de mama podían recordar sus exposiciones pasadas de modo distinto a las mujeres sin cáncer, lo que hoy conocemos como el sesgo de memoria (o de recuerdo).

En los años siguientes, Lombart y Doehring, en 1928, y Stocks y Karn, en 1933, publicaron sendos estudios que hoy podríamos calificar como de casos y controles emparejados.

En 1928, Major Greenwood se convirtió en el primer catedrático de Epidemiología en el Reino Unido. En 1935 publicaría el que algunos consideran el primer libro de texto de la epidemiología como disciplina científica, *Epidemics and crowd-diseases*, aunque esa es una afirmación discutible. (22)

Durante este periodo se diseñaron algunos ensayos controlados y aleatorizados, pero aún de modo irregular. El *Therapeutic Trial Committee of the Medical Research Council*, publicado en 1934, fue un ensayo para evaluar la eficacia de la seroterapia en el tratamiento de la neumonía. Se comparaba la administración de un suero rico en anticuerpos contra la neumonía frente al reposo físico. La forma de asignación fue alternante, que muchos pensaban que era la mejor forma de conseguir grupos comparables entre sí. Sin embargo, con el método alternante los clínicos conocían perfectamente a quién le tocaba el tratamiento novedoso y podían manipular la asignación. De hecho, los pacientes asignados a la seroterapia eran más jóvenes y tenían mejor pronóstico. Esta limitación hizo que en 1941 Joseph Asbury Bell, un discípulo de Frost, dirigiera un ensayo sobre la vacuna frente a la tos ferina con ocultación del reparto aleatorio; en opinión de Iain Chalmers, ese fue el primer ensayo controlado y aleatorizado en el que sus aspectos metodológicos se publicaron con mucho detalle. (23)

Los conceptos básicos de la epidemiología se fueron definiendo durante el periodo de la epidemiología temprana. Hubo un esfuerzo por aumentar la comparabilidad de los grupos, sin mencionar el término confusión, en el desarrollo de

los diseños de cohortes, de casos y controles y de los ensayos controlados y aleatorizados, como hemos visto antes. También se empleó la estandarización de riesgos y de tasas e incluso una ecuación de probabilidad de similitud (*propensity score*) de la exposición.

Por otra parte, Yule, Greenwood y Hill, en el Reino Unido, advirtieron sucesivamente del peligro de agrupar distintos estratos para obtener medidas de asociación o de efecto promedio, pues puede dar lugar a asociaciones espurias inexistentes en el análisis de los estratos por separado. Aunque no fueron más allá.

Al comienzo de este periodo, la epidemiología se definía como el estudio de la historia natural y de las causas de las enfermedades infecciosas. Esta definición se fue ampliando a las enfermedades carenciales y, por último, al estudio de cualquier enfermedad que se manifestara como un fenómeno masivo.

También conviene recordar otros nombres relacionados con la epidemiología social, como fueron los de Florence Nightingale (1820-1910), la fundadora de la enfermería moderna, a cuyo desarrollo contribuyó enormemente, que también tuvo un papel crucial para promover la higiene;⁹ W. E. B. Du Bois (1868-1963), que combatió los prejuicios racistas de la sociedad estadounidense y demostró que las supuestas diferencias sociales de raza eran en gran parte diferencias de clase social; Alice Hamilton (1869-1970), que dedicó su vida profesional al estudio de las enfermedades profesionales, como la intoxicación por plomo o por anilinas, y Salvador Allende (1908-1973), el que fuera presidente de Chile, víctima del golpe de Estado de 1973, que en 1939, cuando era ministro de salubridad, publicó *La realidad médico-social chilena*, en donde calificaba la tuberculosis como una enfermedad social y remarcaba la importancia de la exposición a los factores sociales en la etiología de las enfermedades infecciosas. (24)

2.3. Epidemiología clásica

En 1948 se publicó el que se considera el primer ensayo controlado y aleatorizado de la época actual, el promovido por el Medical Research Council británico para evaluar el efecto de la estreptomycin en el tratamiento de la tuberculosis pulmonar. Fue el primer ensayo en donde se explicó con claridad la ocultación de la asignación aleatoria para garantizar que los grupos de estudio fueran lo más comparables posible, en ese caso mediante sobres sellados. Algunos autores apuntan

⁹ En los cuarteles de Selimiye (Turquía), durante la guerra de Crimea, en 1855, Florence Nightingale promovió medidas higiénicas, como la limpieza de los vertederos y la mejora de la ventilación, que redujeron drásticamente la mortalidad de los soldados británicos. Ella misma recogió los datos de mortalidad rigurosamente y aplicó métodos estadísticos para la vigilancia de los casos. También contribuyó a la reforma hospitalaria y promovió las leyes de salud pública en 1875. (14)

a un ensayo anterior, también promovido por el Medical Research Council, para intentar la ocultación de la asignación, aunque con un método más rudimentario, diseñado por Philip D'Arcy Hart para estudiar el efecto terapéutico de la pitulina en los síntomas del catarro común publicado en 1944. (25,26) Unas décadas más tarde se recopilarían y desarrollarían los métodos y los análisis estadísticos de los ensayos controlados y aleatorizados. (27,28)

La investigación sobre el tabaco presidió esta etapa. Aunque ya había sospechas de la asociación entre el tabaco y el cáncer de pulmón,¹⁰ al comienzo de la década de 1950 se publicaron varias investigaciones ya enteramente epidemiológicas, tres estudios de casos y controles, dos en EE. UU. y uno en el Reino Unido. Este último, firmado por Richard Doll y Austin Bradford Hill, que ellos denominaron un *estudio retrospectivo*, solo usó pruebas de significación estadística y no medidas de efecto.

Doll y Hill iniciaron poco después un estudio de cohortes, el *British Doctor Study*, para evaluar mejor los efectos del tabaco, ante la sospecha de que pudiera haber algún defecto en los estudios de casos y controles. Reclutaron más de 40 000 médicos en total. Desde sus primeras publicaciones, en 1954, se pudo comprobar que las medidas de efecto estimadas por los dos diseños eran muy similares.

En EE. UU. se lanzó en 1952 el *Hammond and Horn Study*, con 200 000 participantes tomados de los miembros de la American Cancer Society y sus familiares y amigos. Demostraron que el tabaco tenía una asociación mayor con el cáncer de pulmón que con la cardiopatía isquémica, pero que esta suponía un exceso de muertes asociadas al tabaco mucho mayor que el cáncer de pulmón. En 1953 se inició otro estudio de cohortes mayor aún, el estudio de los veteranos de EE. UU., que reclutó a 300 000 varones.

Curiosamente, dos bioestadísticos, el británico Ronald Fisher (1890-1962, y artífice de la aleatorización en los experimentos), y el americano Joseph Berkson (1899-1982), cuestionaron que la asociación entre el tabaco y el cáncer de pulmón fuera causal y llegaron a decir que tanto el cáncer de pulmón como el hábito de fumar podrían deberse a causas comunes. Fisher incluso sugirió que el cáncer de pulmón podría ser la causa que indujera a fumar. Tanto Fisher como Berkson desconfiaban de los estudios observacionales, pues no podían descartar los sesgos y lo

¹⁰ Aunque apenas tuvo repercusión fuera de Alemania, en 1944 Schairer y Schöniger publicaron un estudio de casos y controles de la asociación del tabaco con el cáncer de pulmón, financiado por las instituciones nazis. Este estudio fue muy avanzado metodológicamente, con dos tipos de controles, uno poblacional y otro hospitalario, y en el que se tuvieron en cuenta los cambios en el hábito de fumar después de la enfermedad. El cálculo de la *odds ratio* del riesgo de cáncer de pulmón en los fumadores empedernidos de ese estudio, 16,6, fue prácticamente igual que el calculado a partir del estudio de casos y controles publicado por Doll y Hill en 1952. (29-31)

que después se llamaría confusión. Otros bioestadísticos se sumaron al escepticismo por dos razones: que el tabaco pudiera ocasionar muchas otras enfermedades contradiría la especificidad conocida de los agentes infecciosos, que solo producían enfermedades individuales; y faltaba el agente cancerígeno concreto, ya que para entonces los estudios de toxicidad animal con el tabaco habían sido infructuosos.¹¹

En 1951, Jerome Cornfield (1912-1979) demostró que en los estudios de casos y controles la razón de riesgos podría estimarse mediante la *odds ratio*, la razón de las posibilidades (*odds*) de exposición en los casos dividida entre las posibilidades de exposición en los controles. Esta aproximación se basaba en una condición, la después denominada «suposición de enfermedad rara». Esta condición se podía aplicar sin problemas al caso del tabaco y el cáncer de pulmón porque la prevalencia del cáncer de pulmón en la población general es de 15,5 por 100 000. Cornfield añadió otra segunda condición, que los casos y los controles fueran representativos de todos los casos y de todas las personas sin enfermedad en la población diana. (32)

El concepto de confusión se desarrolló cuando el estadístico británico Edward H. Simpson describió lo que se conoce como la paradoja que lleva su nombre en 1951. Simpson estableció la definición clásica de variable confusora, una tercera variable distinta a las de exposición y de desenlace que se relacione con el desenlace en los no expuestos y con la exposición en los que no presentan el desenlace. En 1959, Nathan Mantel (1919-2002) y William Haenszel (1910-1998) publicaron un artículo clásico con la estimación de la *odds ratio* ajustada, cuando se busca una *odds ratio* promedio de varios estratos de la variable de confusión. Aunque no emplearon los términos *odds ratio* ni confusión. (33)¹²

El término confusión apareció en la literatura epidemiológica a partir de los años 70, como recogen los libros de texto de MacMahon y Pugh y de Susser. Se ilustraban con ejemplos en donde la *odds ratio* o las razones de riesgos eran homogéneas en todos los estratos de la variable de confusión, pero la medida de efecto cuando se agrupaban todos los casos era distinta.

¹¹ Aunque ahora nadie discute la relación causal entre el tabaco y el cáncer de pulmón, en su día esta actitud discrepante resultó ser muy enriquecedora, pues contribuyó al desarrollo de nuevos estudios que confirmaran esa causalidad y a mejorar los métodos de los estudios observacionales analíticos. El escepticismo en la ciencia es saludable, siempre que se base en argumentos razonables. Esto no tiene nada que ver con las dudas maliciosas inducidas por intereses creados, como los de la industria tabacalera para cuestionar los efectos del tabaco sobre la salud o los de la industria de los combustibles fósiles para negar su responsabilidad en la génesis del calentamiento global. Para que todas las cartas estén sobre la mesa, como mínimo todos los investigadores deben declarar sus conflictos de intereses.

¹² El término empleado antes de que apareciera *odds ratio* era riesgo relativo, tanto en los trabajos de Cornfield como en el de Mantel y Haenszel. La primera mención al término *odds ratio* en Medline aparece en 1970 en un trabajo de Joseph L. Fleiss (1937-2003), uno de los padres de la bioestadística moderna. (34)

El concepto de interacción lo desarrollaron Hammond *et al.* en un artículo publicado en 1979. (35) Encontraron que tanto el tabaco como el amianto producían muertes por cáncer de pulmón, pero que la diferencia de riesgos de la exposición conjunta con respecto al grupo control era mucho mayor que la suma de las diferencias de riesgos individuales de cada exposición, es decir, había un efecto sinérgico en la escala aditiva.

El concepto de sesgo de selección en la teoría clásica se atribuye a Joseph Berkson, que trabajaba en la Clínica Mayo. Berkson se sorprendió al leer un estudio de casos y controles publicado en 1929 por Raymond Pearl, biólogo y estadístico de la Johns Hopkins University, en el que analizaba los resultados de las autopsias de 816 pacientes fallecidos por cáncer y los comparaban con los de 816 personas emparejadas por sexo, edad y raza que habían muerto por otras causas. En ese estudio, Pearl encontró que en los casos con cáncer había un 6,6 % de fallecidos con tuberculosis, frente al 16,3 % en los controles. Se encontraba una asociación inversa aparente entre la tuberculosis y el cáncer. En 1946 publicó un artículo en el que creía demostrar que dos enfermedades que eran independientes en la población general podían asociarse espuriamente en estudios de casos y controles hospitalarios, lo que después se llamaría la falacia, la paradoja o el sesgo de Berkson, según quienes fueran los que la denominaran. (36)

El concepto de errores de clasificación (sesgo de información) también se desarrolló en los años 50. Ernst Wynder (1922-1999) y coautores publicaron en 1954 un artículo en el que describían que la circuncisión del marido protegía a su esposa del cáncer de cuello uterino. (37) En 1958 Lilienfeld y Graham cuestionaron que las esposas fueran una fuente válida para saber si sus maridos se habían circuncidado. Estos autores compararon el diagnóstico de circuncisión hecho por un médico que exploró a los hombres con la autorreferida por ellos y encontró un 56 % de falsos negativos y un 18 % de falsos positivos en las autodescripciones. (15) Se generó una polémica sobre el impacto de la clasificación errónea de la exposición sobre el efecto, polémica que finalmente se aclaró con los trabajos de Newell (38) y de Copeland *et al.* (39)

Durante esta etapa proliferaron las descripciones de posibles sesgos de todo tipo en los estudios epidemiológicos. David L. Sackett (1934-2015) publicó en 1979 un intento de catalogarlos, en el que identificó 35 fuentes de sesgos y los dividió según aparecen en el diseño, en la ejecución, en la medición, en el análisis o en la interpretación de un estudio. (40)

En 1965, Austin Bradford Hill publicó una serie de características para juzgar las relaciones causales en la epidemiología. (41) Esas características se convirtieron en elementos decisivos para la reflexión sobre la inferencia causal que hasta ahora han resistido el paso del tiempo, a pesar de que por sí mismas no pueden sentar

una relación causal y a pesar de la crítica a la que las han sometido desde entonces. En el fondo subyacía la cuestión, en muchas ocasiones no reconocida, de cuál es el grado de certidumbre necesario para establecer relaciones causales de determinadas exposiciones con algunas enfermedades cuando se trata de tomar decisiones de salud pública.

En resumen, esta etapa de la epidemiología clásica fue la de la consolidación de la epidemiología como disciplina científica, con el desarrollo de sus conceptos centrales. Se publicaron multitud de libros de texto desde 1956 hasta 1976. Abarca, aproximadamente, entre 1945 y finales de los años 70. Uno de los logros más simbólicos de la contribución de la epidemiología durante este periodo fue la publicación en 1964 del informe del Cirujano General de EE. UU., *Smoking and Health*, el trabajo de expertos que reunió las pruebas epidemiológicas y concluyó que había una asociación muy poderosa entre el tabaco de fumar y el cáncer de pulmón.¹³

2.4. Epidemiología moderna

En 1976 Olli S. Miettinen publicó el artículo *Estimability and estimation in case-referent studies* que se considera como el que marca la transición de la etapa de la epidemiología clásica a la epidemiología moderna. (43) En ese artículo establece la distinción y la relación entre el riesgo y la tasa (denominados entonces incidencia acumulada y la densidad de incidencia) y el vínculo entre los estudios de casos y controles y los estudios de cohortes. Además, Miettinen cuestionó que la *odds ratio* solo pudiera aproximarse a la razón de riesgos cuando se cumplía la suposición de enfermedad rara. Eso dependía del modo como se seleccionasen los controles

¹³ El Cirujano General (*Surgeon General*) de EE. UU. es el jefe del Cuerpo Comisionado del Servicio de Salud Pública de los Estados Unidos y, como tal, es el portavoz en materia de salud pública del Gobierno federal del país. Lo nombra el presidente del país con la confirmación del Senado.

El informe *Smoking and Health* fue preparado por un comité de diez miembros, dos de ellos consultores de la industria tabacalera y el resto aceptados por ella, que tenían derecho de veto. El Cirujano General también vetó a todos aquellos que se habían pronunciado públicamente sobre los efectos del tabaco sobre la salud. En sus resultados, tras insistir en que los métodos estadísticos no pueden probar una relación causal cuando hay una asociación, el informe decía que «a menudo, es necesaria la coexistencia de varios factores para que se produzca una enfermedad, y que uno de esos factores puede jugar un papel determinante; es decir, sin ese factor el resto de factores (como una susceptibilidad genética) apenas conducen a la aparición de la enfermedad», lo que es cierto. A pesar de que el informe fue fruto del consenso de sus miembros y, por tanto, no hubo opiniones discrepantes, y a pesar de la composición del comité, la industria tabacalera lo recibió con hostilidad e insistió en que la asociación era meramente estadística, no causal, y promovió a expertos bien pagados que lanzaron la idea equivocada de que la comunidad científica discrepaba sobre si el tabaco era causa del cáncer de pulmón y, sobre todo, de otras patologías. La industria también explotó a su favor que el informe no reconocía el consumo de tabaco como adicción —lo que no se reconoció oficialmente hasta el informe del Cirujano General de 1988, 24 años más tarde—, sino como un mero hábito. (42)

a partir de las cohortes hipotéticas subyacentes. El supuesto de enfermedad rara era adecuado cuando los controles se seleccionaban entre los individuos que no habían desarrollado el desenlace al final del periodo en riesgo. Pero ese supuesto no era necesario cuando los controles eran una muestra del conjunto de la cohorte inicial (lo que ahora llamamos estudio de cohorte y casos, en donde la *odds ratio* estima la razón de riesgos) o cuando estaban todavía en riesgo de tener el desenlace en el momento en que se seleccionaba el caso (como sucede en los estudios de casos y controles anidados en una cohorte, en donde la *odds ratio* estima la razón de tasas o de densidades de incidencia).

Los conceptos de confusión, interacción y sesgos se han ido refinando. A la definición de variable de confusión formulada por Simpson (que se asocia con la exposición en la población y con el desenlace en los no expuestos) al final de los años 70 y en la década de los 80 se añadió una tercera condición, que la variable no fuera un mediador en la vía causal entre la exposición y el desenlace. Pero una variable puede cumplir esas tres condiciones sin ser una variable de confusión. Tampoco sirve para definir como confusión la circunstancia en que las medidas de asociación o de efecto de los estratos de la variable confusora no coinciden con las medidas brutas no estratificadas, lo que se conoce como la *no colapsabilidad*, pues causas distintas a la confusión pueden explicar esa discrepancia. En 1986, Greenland y Robins abordaron la confusión desde un enfoque de desenlaces potenciales en el que la población expuesta ha de ser comparable (o *intercambiable*) con la población no expuesta para poder establecer una relación causal entre la exposición y el desenlace. La confusión es, por tanto, una cuestión de falta de comparabilidad entre los grupos que se estudian. (44)

El concepto de interacción apenas tuvo presencia en la epidemiología antes de la aparición de un artículo de Rothman en 1974. (45) A partir de entonces se ha desarrollado una amplia bibliografía para evaluar, cuantificar y comprobar las interacciones en las escalas aditivas y multiplicativas. El concepto de interacción resulta indispensable para comprender el modelo causal de causas suficientes que describió Rothman. (46) Llevado al extremo, si uno de los componentes de una causa suficiente está ausente, no se produce el desenlace. Pero el desenlace sí se produce cuando todos los componentes de la causa suficiente están presentes.

En este periodo, la epidemiología ha clasificado los sesgos en dos tipos, los de selección y los de información, también llamados de medición o de clasificación errónea.¹⁴ Fue Copeland quien analizó los sesgos de clasificación errónea y estable-

¹⁴ En la literatura epidemiológica hay dos escuelas discrepantes sobre si la confusión se concibe como sesgo o no. En este capítulo se opta por considerar la confusión como un fenómeno distinto de los sesgos. Una asociación confundida es una asociación real, que existe, pero que no es causal, sino meramente estadística. Cuando hay sesgos, la asociación encontrada no es real, es falsa. (47,48)

ció la regla de que esos sesgos tendían a subestimar el efecto real cuando la clasificación errónea de la exposición era similar (no diferencial) en los grupos definidos por la variable de desenlace o viceversa. (39)

El sesgo de selección, sobre el que había cierta incoherencia en su definición al comienzo del periodo, quedó definido como la distorsión de la medida de efecto producida cuando la exposición y el desenlace tienen un efecto común. No obstante, y aunque no sea lo habitual, también puede producirse un sesgo de selección en ausencia de efectos comunes de la exposición y el desenlace (ver capítulo de gráficos causales). En cambio, la confusión se producía cuando la exposición y el desenlace tenían una causa común. (49)

En 1992 se publicó el que podría calificarse como el documento fundacional de la mal traducida medicina basada en la evidencia, es decir, la medicina basada en pruebas científicas: la consolidación del cuerpo de conocimientos y de los métodos de la epidemiología para generar pruebas científicas sólidas que sirvan de base para una práctica clínica mejor, junto con la experiencia de los profesionales de salud y las preferencias de los pacientes. (50) Esto supuso la confirmación del papel central que la epidemiología, como disciplina científica, tiene en todos los aspectos de la clínica, como la etiología, la prevención, el diagnóstico, el pronóstico, el tratamiento y la rehabilitación. Este movimiento es la culminación de los esfuerzos promovidos desde los años 80 para establecer la epidemiología clínica como una entidad propia, por el grupo de Sackett (51), y por otros autores, como Alvan Feinstein, Robert H. Fletcher, Suzanne W. Fletcher, Edward H. Wagner y Noel Weiss. El documento fundacional de la medicina basada en pruebas fue liderado desde la Universidad de McMaster, en Hamilton, Ontario, por David L. Sackett y, posteriormente, por Gordon H. Guyatt —quien acuñó el término *evidence based medicine* en 1991 y es uno de los promotores del enfoque GRADE (*Grading of Recommendations Assessment, Development and Evaluation*) para la evaluación de la calidad de la evidencia y de la fuerza de las recomendaciones de las intervenciones sanitarias—. (52)¹⁵

La medicina basada en pruebas impulsó el desarrollo de las revisiones sistemáticas y al metanálisis. Archie Cochrane (1909-1988), un médico escocés, fue el inspirador de las revisiones sistemáticas. En su libro de 1972 *Effectiveness and efficiency. Random*

¹⁵ La medicina basada en las pruebas científicas, un término que no deja de ser una obviedad, ha sido objeto de fuertes críticas, algunas justificadas. Unas críticas provienen del campo de la filosofía, en particular de la filosofía de la ciencia, por varios motivos. Entre otros, por la indefinición del concepto de *evidencia* o de pruebas científicas, por suponer que la objetividad científica carece de valores y no está inserta en el contexto social de la práctica médica de su tiempo, es decir, por pensar que los hechos son ajenos a los valores y por no considerar suficientemente la incertidumbre inherente a los pacientes individuales cuando se les aplica acciones tomadas de estudios en grupos. (53-55) Otras críticas dirigen sus dardos a la monopolización por la industria farmacéutica de la investigación relacionada con la salud y en particular de los ensayos controlados y aleatorizados, lo que ha marcado su agenda en múltiples planos, y a las consecuencias no deseadas de su aplicación mecánica e inflexible. (56,57)

reflections on health services lamentaba la falta de evidencia científica, de pruebas sólidas, en muchas de las intervenciones sanitarias y el uso ineficiente de los recursos, y señaló la importancia de los ensayos controlados y aleatorizados para evaluar la efectividad de los tratamientos. (58) Aunque el término *metanálisis* fue acuñado por el estadístico Gene V. Glass en 1976, su vigencia en la literatura médica se desarrolló en los años 80 y 90, (59,60) con trabajos colaborativos sobre el tratamiento del infarto agudo de miocardio —con aspirina, betabloqueantes y estreptoquinasa—, el cáncer de mama y la hipertensión arterial, casi siempre con ensayos clínicos, aunque también con estudios observacionales, y en ocasiones con datos individuales. La Colaboración Cochrane fue fundada en 1993 por Iain Chalmers y Peter C. Gøtzsche siguiendo el espíritu de Archie Cochrane para evaluar las intervenciones sanitarias mediante revisiones sistemáticas.

En este periodo, la epidemiología se ha orientado al estudio de múltiples factores de riesgo de las llamadas enfermedades crónicas. La epidemiología moderna ha unificado los distintos componentes de la teoría de enfermedades transmisibles y no transmisibles relativos a las medidas de frecuencia de la enfermedad, el diseño de los estudios, la confusión, la interacción y los sesgos.

Durante esta fase, se han publicado muchos libros de texto, como la serie de los liderados por Rothman con el mismo título de *Epidemiología moderna*, (61-64) las seis ediciones del libro de Leon Gordis, desde 1996 hasta 2019, (47) y las cuatro ediciones del libro de Szklo y Nieto, desde 2000 hasta 2019. (48) En estos libros se concede una importancia relevante a la aplicación de los métodos epidemiológicos a campos específicos, como el medioambiente, la nutrición, la genética o la clínica.

En algunos libros de texto, la definición de epidemiología ha perdido su segundo componente, el de «la aplicación de este conocimiento para el control de problemas de salud relevantes», de modo que se desvanece su relación práctica con la salud pública, (62) quizá por ver en ese aspecto práctico algo que se da por supuesto, al menos en los países ricos, o quizá por entender que eso concierne a otros. Sin embargo, la epidemiología cobra sentido como disciplina científica al servicio de la salud pública, por lo que cabe pensar que la definición clásica, con los dos componentes, teórico y práctico, prevalezca en el futuro. (47)

Anthony J. McMichael (1942-2014) fue un epidemiólogo australiano que estudió los efectos nocivos de la exposición al plomo de las mujeres embarazadas sobre el desarrollo intelectual a largo plazo de sus hijos, y también fue un pionero en el estudio de los efectos sobre la salud del cambio climático y de otras amenazas medioambientales, por lo que llegó a dirigir la redacción de varios informes sobre la salud del Grupo Intergubernamental de Expertos sobre Cambio Climático (*Intergovernmental Panel on Climate Change*, IPCC). Criticó la epidemiología moderna por estar «prisionera de lo próximo» en cuatro facetas: el énfasis en los factores

de riesgo individuales y próximos no se acompaña de una preocupación por lo que sucede río arriba, por los fenómenos sociales y poblacionales;¹⁶ el análisis causal se ciñe al nivel individual sin tener en cuenta los niveles grupales o poblacionales; se pone el foco en las exposiciones que ocurren en un estado presente o próximo en el pasado pero no se consideran los modelos de riesgo que suceden a lo largo de la vida, y, por último, no se anticipan los riesgos para la salud que provienen de los cambios ecológicos globales a los que cada vez estamos más expuestos. (66)

McMichael criticaba la epidemiología moderna no por lo que hace, que ha sido muy productivo, aunque con rentabilidades decrecientes, sino porque solo hace eso y no aborda los planos señalados. Su crítica se basaba, en parte, en la de Mervyn Susser (1921-2014), que fue un epidemiólogo sudafricano que desarrolló en EE. UU. la mayor parte de su carrera. Susser hizo contribuciones muy valiosas a la epidemiología en muchas facetas, entre ellas el estudio de la causalidad y un desarrollo de la epidemiología social. Susser defendió lo que llamó la ecoepidemiología, que aborda todos los planos relacionados con el estudio de la enfermedad, desde el micro o molecular hasta el macro, o ecológico o social, pasando por el personal o conductual de los factores de riesgo individuales. (67) Junto a Susser, otros autores como Geoffrey Rose (1926-1993), Nancy Krieger, Sally Zierler y Neil Pearce también pusieron el énfasis en las causas sociales de la enfermedad desde una perspectiva poblacional. (7,36,68,69)¹⁷

Ahora nos encontramos en una fase de transición de la epidemiología, una vez más. Cómo vaya a ser la nueva etapa no lo podemos saber aún, pero llegará un nuevo paradigma que sustituya al anterior y que esté mejor provisto para estudiar los problemas relacionados con la salud de los seres humanos y aplicar ese conocimiento para controlar esos problemas y mejorar la salud de la población.

¹⁶ El aislamiento de la asociación entre una exposición y la enfermedad de su contexto histórico, social y político ha dado lugar a algunos sonoros fracasos en la investigación epidemiológica, como cuando los estudios observacionales establecieron en las mujeres posmenopáusicas una asociación preventiva del tratamiento hormonal sustitutivo en la cardiopatía isquémica, hasta que el ensayo *Women's Health Initiative* demostró lo contrario. En los estudios observacionales, los grupos que se comparaban entre sí eran muy diferentes en múltiples facetas, lo que expresa la complejidad de las vidas y los tiempos de las personas, inasible al control de la confusión por el ajuste de unas pocas variables; el resultado fue que los grupos de la exposición y los grupos de control siguieran siendo muy distintos entre sí con respecto a muchos aspectos que influían en el desenlace. (31)

¹⁷ Neal Pearce resaltaba que, a pesar de las contribuciones de la epidemiología en las últimas décadas, los efectos nocivos del tabaco y del asbesto no se han reducido en el mundo, solo se han desplazado hacia los países con menos ingresos. En su opinión, la epidemiología se había convertido en un conjunto de métodos, de modo que los epidemiólogos intentaban estudiar solo las preguntas de investigación que podían abordarse con esos métodos. Eso llevaba a despreciar los estudios ecológicos, sin los cuales muchos problemas no pueden estudiarse, (70) En las dos últimas décadas la situación ha cambiado un poco, y la contribución de los estudios ecológicos al conocimiento de algunos fenómenos, como los efectos sobre la salud de la contaminación atmosférica, ha influido en ello. Pearce también ponía el foco en la influencia de las corporaciones privadas (en múltiples ámbitos, como el tabaco, los medicamentos y dispositivos médicos, las sustancias químicas...) en la investigación epidemiológica. (70) Eso ha cambiado poco.

3. Referencias bibliográficas

1. Porta M, Greenland S, Hernán M, dos Santos Silva I, Last JM, editores. A dictionary of epidemiology. Sexta edición. Oxford: Oxford University Press; 2014.
2. Corominas J. Breve diccionario etimológico de la lengua castellana. Tercera edición. Madrid: Editorial Gredos; 1973. Página 204.
3. Cole SR. Atheoretical science: a response to the poverty of Popperian epidemiology. *Int J Epidemiol*. 1996;25:899-900. doi: [10.1093/ije/25.4.899](https://doi.org/10.1093/ije/25.4.899).
4. Susser M. Causal thinking in the health sciences: concepts and strategies of epidemiology. New York: Oxford University Press; 1973.
5. Last JM. A dictionary of epidemiology. Segunda edición. New York: Oxford University Press; 1988.
6. Pearce N. A short introduction to epidemiology. Segunda edición. Wellington, Nueva Zelanda: Centre for Public Health Research, Massey University Wellington Campus; 2005. https://vula.uct.ac.za/access/content/group/9c29ba04-b1ee-49b9-8c85-9a468b556ce2/DOH/Module%202%20%28Bio_Epi%29/Epidemiology/EPIDEMIOLOGY/Pearce.pdf. Consultado el 3 de octubre de 2023.
7. Rose G. The strategy of preventive medicine. Oxford: Oxford University Press; 1992.
8. López-Moreno S, Garrido-Latorre F, Hernández-Avila M. Desarrollo histórico de la epidemiología: su formación como disciplina científica. *Salud Pública Mex*. 2000;42:133-43. doi: [10.1590/S0036-36342000000200009](https://doi.org/10.1590/S0036-36342000000200009).
9. Kayali G. The forgotten history of pre-modern epidemiology: contribution of Ibn An-Nafis in the Islamic golden era. *East Mediterr Health J*. 2018;23:854-7. doi: [10.26719/2017.23.12.854](https://doi.org/10.26719/2017.23.12.854).
10. Tuells J. Early mention of the term epidemiology. *Emerg Infect Dis*. 2016;22:2006. doi: [10.3201/eid2211.141466](https://doi.org/10.3201/eid2211.141466).
11. Buck C, Llopis Á, Nájera E, Terris M, editores. El desafío de la epidemiología. Problemas y lecturas seleccionadas. Washington, D.C.: Organización Panamericana de la Salud; 1988. <https://iris.paho.org/handle/10665.2/34153>. Consultado el 3 de octubre de 2023.
12. Morabia A. Pandemics and methodological developments in epidemiology history. *J Clin Epidemiol*. 2020;125:164-9. doi: [10.1016/j.jclinepi.2020.06.008](https://doi.org/10.1016/j.jclinepi.2020.06.008).

13. Morabia A. History of epidemiological methods and concepts. En: Ahrens W, Pigeot I, editors. Handbook of epidemiology. Segunda edición. New York: Springer; 2014. Páginas 43-74.
14. Merrill RM. Historic developments in epidemiology. En Merrill RM. Introduction to epidemiology. Séptima edición. Burlington, MA: Jones & Bartlett Learning LLC; 2016. Páginas 17-37.
15. Lilienfeld AM, Lilienfeld DE. Foundations of epidemiology. Segunda edición. New York: Oxford University Press; 1980.
16. Dubois JC. Epidemic diseases and Chinese medicine: example of SARS and COVID-19: part two. *Chinese Medicine and Culture*. 2020;3:121-32. doi: [10.4103/CMAC.CMAC_32_20](https://doi.org/10.4103/CMAC.CMAC_32_20).
17. Rothman KJ. The rise and fall of epidemiology, 1950--2000 A.D. *N Engl J Med*. 1981;304:600-2. doi: [10.1056/NEJM198103053041010](https://doi.org/10.1056/NEJM198103053041010).
18. Jamrozik E, Selgelid MJ. COVID-19 human challenge studies: ethical issues. *Lancet Infect Dis*. 2020;20:e198-e203. doi: [10.1016/S1473-3099\(20\)30438-2](https://doi.org/10.1016/S1473-3099(20)30438-2).
19. Miller FG, Moreno JD. Human infection challenge experiments: then and now. *Ethics Hum Res*. 2021;43:42-4. doi: [10.1002/eahr.500088](https://doi.org/10.1002/eahr.500088).
20. Morabia A. Pierre-Charles-Alexandre Louis and the evaluation of bloodletting. *J R Soc Med*. 2006;99:158-60. doi: [10.1177/014107680609900322](https://doi.org/10.1177/014107680609900322).
21. Vandenbroucke JP, Eelkman Rooda HM, Beukers H. Who made John Snow a hero? *Am J Epidemiol*. 1991;133:967-73. doi: [10.1093/oxford-journals.aje.a115816](https://doi.org/10.1093/oxford-journals.aje.a115816).
22. Bracken MB. The first epidemiologic text. *American Journal of Epidemiology*. 2003;157:855-6. doi: [10.1093/aje/kwg060](https://doi.org/10.1093/aje/kwg060).
23. Chalmers I. Joseph Asbury Bell and the birth of randomized trials. *J R Soc Med*. 2007;100:287-93. doi: [10.1177/014107680710000616](https://doi.org/10.1177/014107680710000616).
24. Foster JB, Clark B, Holleman H. Capital and the ecology of disease. Monthly Review. Junio 2021. <https://monthlyreview.org/2021/06/01/capital-and-the-ecology-of-disease/>. Consultado el 3 de octubre de 2023.
25. Streptomycin treatment of pulmonary tuberculosis. A Medical Research Council Investigation. *Br Med J*. 1948;2:769-82. doi: [10.1136/bmj.2.4582.769](https://doi.org/10.1136/bmj.2.4582.769).
26. Chalmers I. Why the 1948 MRC trial of streptomycin used treatment allocation based on random numbers. *J R Soc Med*. 2011;104:383-6. doi: [10.1258/jrsm.2011.11k023](https://doi.org/10.1258/jrsm.2011.11k023).

27. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer*. 1976;34:585-612. doi: [10.1038/bjc.1976.220](https://doi.org/10.1038/bjc.1976.220).
28. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *Br J Cancer*. 1977;35:1-39. doi: [10.1038/bjc.1977.1](https://doi.org/10.1038/bjc.1977.1).
29. Schairer E, Schöniger E. Lung cancer and tobacco consumption. *Int J Epidemiol*. 2001;30:24-7. doi: [10.1093/ije/30.1.24](https://doi.org/10.1093/ije/30.1.24).
30. Davey Smith G, Ströbele S, Egger M. Smoking and death. Public health measures were taken more than 40 years ago. *BMJ*. 1995;310:396. doi: [10.1136/bmj.310.6976.396](https://doi.org/10.1136/bmj.310.6976.396).
31. Davey Smith G. Reflections on the limitations to epidemiology. *J Clin Epidemiol*. 2001;54:325-31. doi: [10.1016/s0895-4356\(00\)00334-6](https://doi.org/10.1016/s0895-4356(00)00334-6).
32. Cornfield J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *J Natl Cancer Inst*. 1951;11:1269-75. doi: [10.1093/jnci/11.6.1269](https://doi.org/10.1093/jnci/11.6.1269).
33. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22:719-48. doi: [10.1093/jnci/22.4.719](https://doi.org/10.1093/jnci/22.4.719).
34. Fleiss JL. On the asserted invariance of the odds ratio. *Br J Prev Soc Med*. 1970;24:45-6. doi: [10.1136/jech.24.1.45](https://doi.org/10.1136/jech.24.1.45).
35. Hammond EC, Selikoff IJ, Seidman H. Asbestos exposure, cigarette smoking and death rates. *Ann N Y Acad Sci*. 1979;330:473-90. doi: [10.1111/j.1749-6632.1979.tb18749.x](https://doi.org/10.1111/j.1749-6632.1979.tb18749.x).
36. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics*. 1946;2:47-53. (Publicado también en *Int J Epidemiol* 2014;43:511-5, doi: [10.1093/ije/dyu022](https://doi.org/10.1093/ije/dyu022)).
37. Wynder EL, Cornfield J, Schroff PD, Doraiswami KR. A study of environmental factors in carcinoma of the cervix. *Am J Obstet Gynecol*. 1954;68:1016-47. doi: [10.1016/s0002-9378\(16\)38400-9](https://doi.org/10.1016/s0002-9378(16)38400-9).
38. Newell DJ. Errors in the interpretation of errors in epidemiology. *Am J Public Health Nations Health*. 1962;52:1925-8. doi: [10.2105/ajph.52.11.1925](https://doi.org/10.2105/ajph.52.11.1925).
39. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105:488-95. doi: [10.1093/oxfordjournals.aje.a112408](https://doi.org/10.1093/oxfordjournals.aje.a112408).

40. Sackett DL. Bias in analytic research. *J Chronic Dis*. 1979;32:51-63. doi: [10.1016/0021-9681\(79\)90012-2](https://doi.org/10.1016/0021-9681(79)90012-2).
41. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295-300. doi: [10.1177/003591576505800503](https://doi.org/10.1177/003591576505800503).
42. Hall W. The 1964 US Surgeon General's report on smoking and health. *Addiction*. 2022;117:3170-5. doi: [10.1111/add.16007](https://doi.org/10.1111/add.16007).
43. Miettinen O. Estimability and estimation in case-referent studies. *Am J Epidemiol*. 1976;103:226-35. doi: [10.1093/oxfordjournals.aje.a112220](https://doi.org/10.1093/oxfordjournals.aje.a112220).
44. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15:413-9. doi: [10.1093/ije/15.3.413](https://doi.org/10.1093/ije/15.3.413).
45. Rothman KJ. Synergy and antagonism in cause-effect relationships. *Am J Epidemiol*. 1974;99:385-8. doi: [10.1093/oxfordjournals.aje.a121626](https://doi.org/10.1093/oxfordjournals.aje.a121626).
46. Rothman KJ. Causes. *Am J Epidemiol*. 1976;104:587-92. doi: [10.1093/oxfordjournals.aje.a112335](https://doi.org/10.1093/oxfordjournals.aje.a112335).
47. Celentano DD, Szklo M. Gordis Epidemiology. Sexta edición: Elsevier; 2019. Tiene traducción al español: Celentano DD y Szklo M (autores) Gordis Epidemiología, 6ª edición. Barcelona: Elsevier Saunders. Septiembre 2019.
48. Szklo M, Nieto J. Epidemiology. Beyond the Basics. Cuarta edición. Burlington, MA: Jones & Bartlett Learning; 2019.
49. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615-25. doi: [10.1097/01.ede.0000135174.63482.43](https://doi.org/10.1097/01.ede.0000135174.63482.43).
50. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*. 1992;268:2420-5. doi: [10.1001/jama.1992.03490170092032](https://doi.org/10.1001/jama.1992.03490170092032).
51. Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. Boston: Little Brown; 1985.
52. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924-6. doi: [10.1136/bmj.39489.470347.AD](https://doi.org/10.1136/bmj.39489.470347.AD).
53. Goldenberg MJ. On evidence and evidence-based medicine: lessons from the philosophy of science. *Soc Sci Med*. 2006;62:2621-32. doi: [10.1016/j.socscimed.2005.11.031](https://doi.org/10.1016/j.socscimed.2005.11.031).

54. Worrall J. Evidence: philosophy of science meets medicine. *J Eval Clin Pract.* 2010;16:356-62. doi: [10.1111/j.1365-2753.2010.01400.x](https://doi.org/10.1111/j.1365-2753.2010.01400.x).
55. Upshur RE. A call to integrate ethics and evidence-based medicine. *Virtual Mentor.* 2013;15:86-9. doi: [10.1001/virtualmentor.2013.15.1.oped1-1301](https://doi.org/10.1001/virtualmentor.2013.15.1.oped1-1301).
56. Greenhalgh T, Howick J, Maskrey N. Evidence based medicine: a movement in crisis? *BMJ.* 2014;348:g3725. doi: [10.1136/bmj.g3725](https://doi.org/10.1136/bmj.g3725).
57. Jureidini J, McHenry LB. The illusion of evidence based medicine. *BMJ.* 2022;376:o702. doi: [10.1136/bmj.o702](https://doi.org/10.1136/bmj.o702).
58. Bonfill X. Revisiting the value of Archie Cochrane's contributions. *BMJ Evid Based Med.* 2023;28:73-4. doi: [10.1136/bmjebm-2022-112142](https://doi.org/10.1136/bmjebm-2022-112142).
59. Stampfer MJ, Goldhaber SZ, Yusuf S, Peto R, Hennekens CH. Effect of intravenous streptokinase on acute myocardial infarction: pooled results from randomized trials. *N Engl J Med.* 1982;307:1180-2. doi: [10.1056/NEJM198211043071904](https://doi.org/10.1056/NEJM198211043071904).
60. Peto R. Reflections on the design and analysis of clinical trials and meta-analyses in the 1970s and 1980s. *J RSocMed.* 2019;112:78-80. doi: [10.1177/0141076818824458](https://doi.org/10.1177/0141076818824458).
61. Rothman KJ. Modern epidemiology. Boston: Little Brown and Co.; 1986.
62. Rothman KJ, Greenland S. Modern epidemiology. Segunda edición. Philadelphia: Lippincott Williams & Wilkins; 1998.
63. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Tercera edición. Philadelphia: Lippincott Williams & Wilkins; 2008.
64. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. Modern epidemiology. Cuarta edición. Philadelphia: WoltersKluwer; 2021.
65. Morabia A. Historical reflections on the book Epidemiologia & Saúde: Fundamentos, Métodos, Aplicações. *Cad Saude Publica.* 2013;29:1059-62. doi: [10.1590/S0102-311X2013000600003](https://doi.org/10.1590/S0102-311X2013000600003).
66. McMichael AJ. Prisoners of the proximate: loosening the constraints on epidemiology in an age of change. *Am J Epidemiol.* 1999;149:887-97. doi: [10.1093/oxfordjournals.aje.a009732](https://doi.org/10.1093/oxfordjournals.aje.a009732).
67. Susser M. Does risk factor epidemiology put epidemiology at risk? Peering into the future. *J Epidemiol Community Health.* 1998;52:608-11. doi: [10.1136/jech.52.10.608](https://doi.org/10.1136/jech.52.10.608).

68. Krieger N, Zierler S. What explains the public's health? A call for epidemiologic theory. *Epidemiology*. 1996;7:107-9. doi: [10.1097/00001648-199601000-00020](https://doi.org/10.1097/00001648-199601000-00020).
69. Pearce N. Traditional epidemiology, modern epidemiology, and public health. *Am J Public Health*. 1996;86:678-83. doi: [10.2105/ajph.86.5.678](https://doi.org/10.2105/ajph.86.5.678).
70. Pearce N. Commentary: The rise and rise of corporate epidemiology and the narrowing of epidemiology's vision. *Int J Epidemiol*. 2007;36:713-7. doi: [10.1093/ije/dym152](https://doi.org/10.1093/ije/dym152).

Capítulo 2.

Medidas de frecuencia y de efecto

JAVIER DAMIÁN

1. Medidas de frecuencia de enfermedad

- 1.1. Incidencia acumulada. Riesgo
- 1.2. Tasa de incidencia
- 1.3. Relaciones entre tasa y riesgo
- 1.4. Prevalencia
- 1.5. Relación entre prevalencia e incidencia
- 1.6. Otras medidas

2. Medidas de efecto

- 2.1. Medidas absolutas. Diferencia de tasas y diferencia de riesgos
- 2.2. Medidas relativas. Razón de tasas. Razón de riesgos. Razón de odds
- 2.3. Medida del efecto en estudios de casos y controles
- 2.4. Medida del efecto en estudios de prevalencia (transversales)
- 2.5. Asociación vs. efecto
- 2.6. Proporciones atribuibles a la exposición

3. Referencias

1. Medidas de frecuencia de enfermedad

La medición de casos de enfermedad u otro tipo de episodios en grupos es la base de los estudios epidemiológicos, tanto para investigar sus causas como para valorar su importancia en salud pública. Existen dos grupos de medidas de frecuencia: de incidencia, si los casos registrados son nuevos, recién ocurridos, y de prevalencia, si son casos existentes en una población en un momento concreto. En ambos casos, para tener una idea apropiada de la magnitud del problema, se relaciona el número de casos con la población y tiempo de donde surgieron.

Existen dos tipos importantes de medidas de incidencia: la incidencia acumulada (o proporción de incidencia) y la tasa de incidencia.

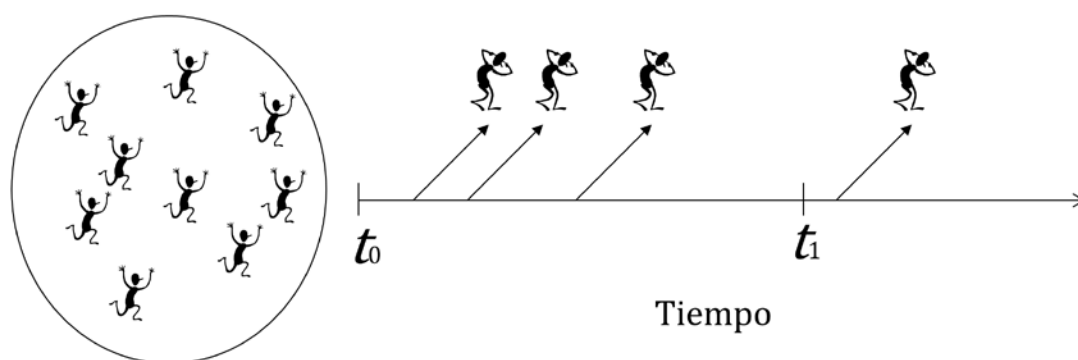
1.1. Incidencia acumulada. Riesgo

La *incidencia acumulada* (*IA*) relaciona el número de casos nuevos durante un periodo de tiempo de observación (t) con el tamaño del grupo al comienzo del mismo:

$$IA = \frac{\text{n.º casos nuevos en el periodo}}{\text{n.º de personas al inicio del periodo}}$$

Esta medida por tanto supone la proporción de personas que han enfermado en ese periodo.

Figura 1. Seguimiento de una cohorte de 9 personas.



Ejemplo 1. En la Figura 1 se observa una población de 9 personas libres de enfermedad en el momento t_0 . Durante el seguimiento se van produciendo episodios de enfermedad. La incidencia acumulada hasta t_1 es de 3 casos nuevos sobre 9 personas al inicio del seguimiento = 0,33.

Esta medida, como toda proporción, no tiene unidades y su rango de posibles valores está entre 0 y 1. Otra característica de las proporciones es que pueden interpretarse en términos de *probabilidad*. La incidencia acumulada por tanto sirve para estimar la probabilidad de desarrollar una enfermedad en un periodo de tiempo, que también se denomina en epidemiología *riesgo*. Cada persona tiene un riesgo diferente de enfermar en un periodo concreto y la incidencia acumulada es el riesgo medio del grupo.

Para interpretar adecuadamente un riesgo es necesario mencionar el periodo de seguimiento, ya que el valor de la incidencia acumulada (probabilidad de ocurrencia de cualquier suceso) tiende a aumentar (tiende a 1) con el paso del tiempo. No tiene sentido informar de una incidencia acumulada de, por ejemplo, 0,12 (o 12 %), sin especificar el tiempo de observación, ya que puede ser muy grande si se refiere a un mes o muy pequeña si se refiere a 30 años.

Una premisa que hay que asumir para estimar adecuadamente el riesgo es que no haya muertes por causa distinta a la enfermedad —o suceso— que se está midiendo. Estas causas se denominan *riesgos competitivos*. Si ocurren no se sabrá si las personas hubieran desarrollado la enfermedad caso de no haber muerto y el riesgo estará infraestimado. Si el riesgo competitivo es pequeño el problema será menor. Por ejemplo, si se hace un seguimiento corto en una población joven la probabilidad de morir es muy baja y los riesgos competitivos para cualquier medición de incidencia serán exiguos.

Otro problema similar ocurre cuando se producen las pérdidas en el seguimiento. Si el periodo de observación es largo la probabilidad de que ocurran estas pérdidas es mayor.

En general se tiene que completar el periodo de observación. Por tanto, salvo cuando se mide el riesgo de morir por cualquier causa, la incidencia acumulada puede entenderse como una estimación de la probabilidad *condicional* de desarrollar la enfermedad en un tiempo determinado.

Además es conveniente calcular el riesgo en una población compuesta por personas que *puedan* desarrollar la enfermedad. Si quisiéramos medir la incidencia acumulada de sarampión en las 9 personas del ejemplo anterior y una tuviera inmunidad deberíamos excluirla del denominador, siendo la incidencia acumulada de la población susceptible $3/8 = 0,37$.

La incidencia acumulada se mide sobre una *población fija o cerrada* que comparte alguna característica inicial (una cohorte). Este tipo de poblaciones se caracterizan por no permitir entradas de nuevos miembros, sólo salidas: porque desarrollan la enfermedad o porque mueren. Sin embargo, en la práctica es habitual que haya personas que se incorporan tarde a la cohorte y que se produzcan pérdidas en

el seguimiento, por lo que los tiempos de observación individuales son diferentes. En las poblaciones fijas la característica o exposición que se estudia es la determinada al comienzo del seguimiento, pero en muchas situaciones las exposiciones cambian. Los diseños sobre poblaciones abiertas o dinámicas permiten solventar parte de los problemas mencionados.

1.2. Tasa de incidencia

Las *poblaciones abiertas o dinámicas* permiten entradas y salidas y cambios de pertenencia a grupos de exposición o características. En estas poblaciones se puede calcular la *tasa de incidencia*, I , basada en denominadores de *persona-tiempo*, que permite analizar datos obtenidos del seguimiento. La tasa de incidencia se define como el número de casos nuevos por cada unidad persona-tiempo de observación:

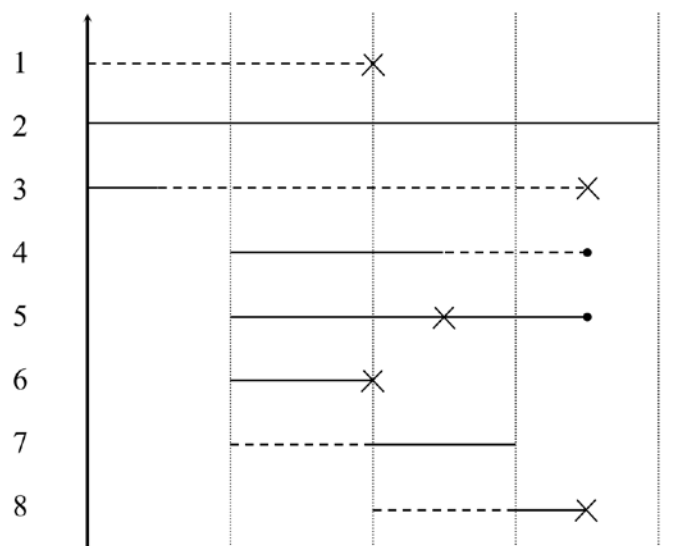
$$I = \frac{\text{n.º de casos nuevos}}{\text{n.º de personas-tiempo en observación}}$$

El número de personas-tiempo se obtiene sumando los tiempos en observación que cada sujeto de estudio está *en riesgo* de ocurrencia de la enfermedad o suceso que se trata de medir. El denominador de la tasa de incidencia está compuesto por tiempo, pero es el tiempo sumado que cada persona aporta, por lo que puede considerarse bidimensional, ya que depende del tamaño del grupo y del tiempo de observación de cada individuo. Es decir, una misma cantidad de personas-tiempo puede ser resultado de seguir a muchas personas poco tiempo o a pocas personas mucho tiempo.

La inclusión de la dimensión temporal en el denominador convierte a la tasa de incidencia en una medida de la *velocidad* con que ocurre una enfermedad en una población.

Al contrario que la incidencia acumulada, que no tiene unidades, la tasa de incidencia sí las tiene, son unidades de tiempo: años, meses, días...; más concretamente del inverso del tiempo, ya que están en el denominador ($1/\text{tiempo}$ o lo que es lo mismo, tiempo^{-1}). Consecuentemente el inverso de la tasa es tiempo ($1/\text{tasa} = \text{tiempo}$) y se interpreta como el tiempo medio hasta la ocurrencia de la enfermedad. El *rango* de valores de la tasa de incidencia está entre 0 e infinito (no tiene un límite superior teórico).

Figura 2. Población dinámica de 8 personas a lo largo de 20 años (líneas verticales separan periodos de 5 años). Las aspas representan casos nuevos de infarto de miocardio. Los puntos negros son muertes. La línea discontinua representa consumo moderado de alcohol. La continua consumo cero.



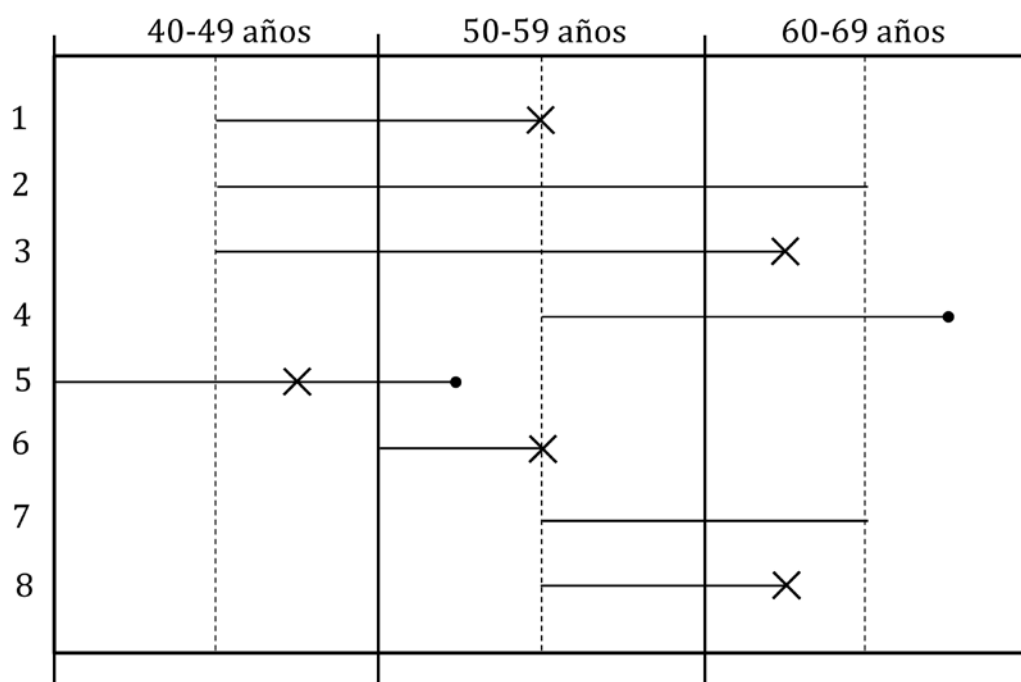
Ejemplo 2. En la Figura 2, se representa el seguimiento durante 20 años a una hipotética población dinámica de 8 personas para medir tasas de incidencia de primeros episodios de infarto de miocardio, representado por un aspa. El punto negro representa los fallecimientos. La cantidad de tiempo de observación que aporta la persona n.º 1 es de 10 años; la persona n.º 2 está en riesgo 20 años. La 3, 17,5 años; la 4, 12,5 años. La persona n.º 5 aporta 7,5 años en riesgo; la 6, 5 años; la 7, 10 años; y la persona n.º 8 aporta 7,5 años. Todo suma 90 años de observación en riesgo, y 5 casos de infarto, por lo que la tasa de incidencia es de $5/90 = 0,056 \text{ año}^{-1}$, o 5,6 casos por cada 100 personas-año.

Al igual que con la incidencia acumulada, durante todo el tiempo que aporta cada persona al denominador de la tasa de incidencia, debe haber tenido la posibilidad de enfermar (tiempo en riesgo). Además debe representar adecuadamente a la población de interés. Por ejemplo, a diferencia de la población cerrada en la que sus características definitorias son las del comienzo del seguimiento, en una población dinámica, una misma persona puede cambiar, y aportar tiempo en riesgo a, o ser parte de, más de una subpoblación diferente. Por ejemplo, cambiando de grupo de edad o de exposición (pasar a ser ex-fumador). En el ejemplo anterior (Figura 2) la parte discontinua representa consumo moderado de alcohol y la parte continua consumo cero. Hay personas que aportan tiempo en riesgo a ambas poblaciones (3, 4, 7 y 8).

1.2.1. Tasa de incidencia específica por edad

Para el calcular las tasas específicas por edad se suman los tiempos de observación que cada persona ha estado en un grupo de edad definido. Esta suma constituye el denominador de la tasa específica en ese grupo de edad y los casos ocurridos en ese grupo forman el numerador.

Figura 3. Las 8 personas de la figura 2 representadas en la escala temporal de la edad. Las aspas representan casos nuevos de infarto de miocardio.



Ejemplo 3. Cálculo de tasas específicas por edad de infarto de miocardio (primeros episodios). La figura 2 muestra la experiencia de 8 personas en la escala de tiempo cronológico. En la figura 3 se representa esa misma experiencia en la escala de edad. La persona 1 entra en el estudio con 45 años justos y sale al cumplir 55, aportando 5 años de tiempo en riesgo (5 personas-año) al grupo de edad de 40-49 años y otros 5 años de tiempo en riesgo y un evento de infarto al grupo de 50-59 años. De esta forma se van calculado personas-año y casos aportados en cada grupo de edad (Tabla 1).

Tabla 1. Tasas de incidencia de infarto de miocardio por edad

Edad	Casos	Personas-año	Tasa (año ⁻¹)	RT _i
40-49	1	22,5	0,044	1,00
50-59	2	45,0	0,044	1,00
60-69	2	22,5	0,089	2,00

En la última columna se muestra la razón entre las tasas (RT_i) de los dos últimos grupos de edad comparados con el primero. Se aprecia una asociación positiva entre el infarto y la edad.

1.2.2. Tasa de incidencia de múltiples episodios

Lo más habitual es que las tasas de incidencia se refieran a primeros episodios de la enfermedad o fenómeno en cuestión. Muchas veces no puede ser de otro modo, como en el caso de mortalidad o de la enfermedad de Parkinson; y en la mayoría de las ocasiones, como el cáncer o la enfermedad cardiovascular, el interés está en la medición de primeras ocurrencias, por lo que el número de casos no puede ser mayor que el número de personas en estudio. Sin embargo, en algunas situaciones se pueden producir múltiples episodios en la misma persona. Un ejemplo son las caídas en personas mayores, que siguen aportando tiempo en riesgo aun después de tener un episodio —siempre y cuando realmente estén en riesgo de que ocurra un nuevo episodio—.

Ejemplo 4. En un estudio la tasa de caídas en personas que viven en residencias para mayores en Madrid fue de 2,4 por cada persona y año. Se produjeron 146 caídas en 733 residentes en un periodo observación de 30 días (1). El 12 % tuvo alguna caída. La tasa de caídas se calculó como

$$\frac{146}{733 \times 30 \text{ días}} = 0,00664 \text{ día}^{-1}$$

lo que supone $0,00664 \text{ día}^{-1} \times 365,25 = 2,4 \text{ año}^{-1}$

En otras situaciones tendrá interés conocer la *tasa de recurrencia*, ya que los fenómenos determinantes de un primer episodio o de una recurrencia pueden ser diferentes. En estos casos la población a riesgo la constituyen personas que ya han tenido un primer episodio y aportan tiempo en riesgo a partir del mismo.

1.2.3. Tasa de incidencia con datos agregados

Cuando no se dispone de la información de seguimiento de cada individuo pueden utilizarse *datos agregados* para calcular tasas de incidencia:

$$I = \frac{\text{Número de casos nuevos en el periodo}}{\text{Población media en el periodo} \times \text{duración del periodo}}$$

La validez de la tasa depende de la calidad de la recogida de datos de incidencia y de la estabilidad del tamaño poblacional por lo que estos cálculos se reservan para poblaciones grandes con buenos registros.

Ejemplo 5. En la tabla 2 se muestra la mortalidad por tumores malignos en España en 2015.

Tabla 2. Mortalidad por tumores en España, 2015

	N.º defunciones ^a	Población a 1 de julio ^b	Tasa ^c
Mujeres	41 632	23 609 285	176
Hombres	65 367	22 800 864	287
Total	106 999	46 410 149	231

^a N.º de defunciones obtenido del Centro Nacional de Epidemiología (Instituto de Salud Carlos III).

^b Instituto Nacional de Estadística de España.

^c Por 100 000 personas-año.

La tasa de mortalidad por tumores malignos en hombres se calcula como

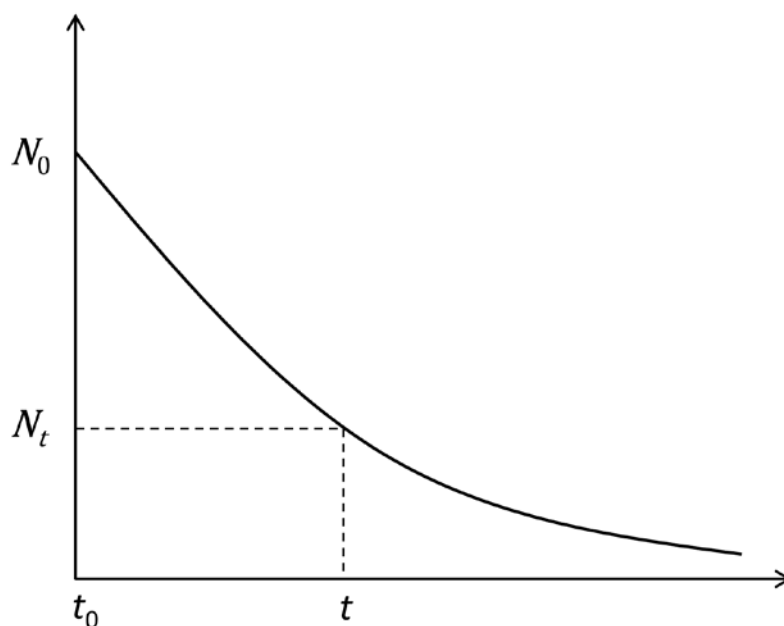
$$I_{\text{hombres}} = \frac{65\,367}{22\,800\,864 \times 1 \text{ año}} = 0,00287 \text{ año}^{-1}$$

Es decir, ocurren 287 muertes por tumores malignos por cada 100 000 hombres y año.

1.3. Relaciones entre tasa y riesgo

La incidencia acumulada se mide sobre una población cerrada. En la Figura 4 se representa la variación en el tamaño de una cohorte con el tiempo. La población inicial, de tamaño N_0 , va enfermando, quedando una población en riesgo de enfermar en un momento t de tamaño N_t . Este número de personas existentes al final de un tiempo de seguimiento de interés depende directamente del número de personas inicial y es función (exponencial decreciente) de la fuerza de morbilidad que produce casos de enfermedad —representada por la tasa de incidencia, I — y del tiempo t :

Figura 4. Reducción en el tamaño de una población fija sometida a una tasa de incidencia constante en función el tiempo.



$$N_t = N_0 \exp(-It)^*$$

$$N_t/N_0 = \exp(-It) \quad [2-1]$$

[*: $\exp(x)$ es otra forma de expresar e^x (el número e : 2,718..., base de los logaritmos naturales)]

Por otra parte, el riesgo de enfermar en un periodo t (de t_0 a t) es:

$$IA = \frac{\text{n.º casos nuevos en el periodo}}{\text{población en } t_0} = \frac{N_0 - N_t}{N_0} = 1 - \frac{N_t}{N_0}$$

por tanto, sustituyendo con [2-1]:

$$IA = 1 - \exp(-It)$$

Para riesgos bajos ($< 0,1$) se puede aproximar a:

$$IA \approx It \quad [2-2]$$

Ejemplo 6. Tomando las tasas de mortalidad por tumores de la tabla 2 se pueden calcular los riesgos en el periodo de tiempo que se elija. Por ejemplo en los hombres el riesgo de mortalidad por cualquier tumor en 10 años es:

$$IA = 1 - \exp(-0,00287 \text{ año}^{-1} \times 10 \text{ años}) = 2,83\%$$

En estas fórmulas se asume que la tasa de incidencia es constante en todo el periodo. Si no se puede asumir esto conviene realizar estimaciones parciales del riesgo en subperiodos (t_i) más cortos. Si se dispone de estimaciones de tasas de incidencia de los subperiodos, la fórmula anterior se generaliza a:

$$IA = 1 - \exp\left(-\sum_i I_i t_i\right)$$

y para riesgos bajos ($< 0,1$):

$$IA \approx \sum_i I_i t_i$$

Ejemplo 7. En el seguimiento realizado a 1 101 669 mujeres durante 19 años se obtuvieron las siguientes tasas de incidencia de cáncer de mama por edad (2):

Tabla 3. Tasas de incidencia de cáncer de mama por edad

Edad	Tasa (año ⁻¹)
25-44	0,00052
45-59	0,00166
60-79	0,00238

El riesgo de desarrollar cáncer de mama en 50 años de una mujer con 30 años será:

$$1 - \exp[-(0,00052 \cdot 15 + 0,00166 \cdot 15 + 0,00238 \cdot 20)] = 8\%$$

1.4. Prevalencia

La prevalencia se define como la *proporción* de sujetos con una determinada característica, normalmente enfermedades o factores de riesgo de las mismas, en un momento o periodo.

$$P = \frac{\text{n.º de personas con la enfermedad}}{\text{total población}}$$

Es una medida estática, como una fotografía donde se refleja la magnitud de un problema en una población en un momento definido. Se denomina *prevalencia puntual* cuando se refiere a un momento concreto y *prevalencia de periodo* cuando incluye las personas que estuvieron enfermas en algún momento de un periodo determinado.

Como es una proporción tiene un rango de valores entre 0 y 1, sin unidades, y puede expresarse en términos de probabilidad, en este caso la probabilidad de que una persona tomada al azar tenga la enfermedad o característica en cuestión.

1.5. Relación entre prevalencia e incidencia

La prevalencia puede considerarse como un subconjunto de la población (Figura 5) compuesto por personas con la enfermedad. A ese subconjunto se entra cuando se enferma y se sale por muerte o curación. El flujo de entradas está determinado por la tasa de incidencia de enfermedad, I . El flujo de salidas está determinado por la tasa de incidencia de “terminaciones de enfermedad”, I^* . Como el inverso de una tasa de incidencia es el tiempo medio hasta la ocurrencia, la tasa de terminaciones es equivalente al inverso del tiempo medio hasta el suceso, que es la terminación de enfermedad, es decir la duración media de la enfermedad: $I^* = 1/\bar{D}$. En una situación estable, sin migraciones, el número de entradas al subconjunto de prevalencia es igual al número salidas. Siendo Cp el n.º de casos prevalentes y N el tamaño de la población:

$$\underbrace{(N - Cp) \times t \times I}_{\text{n.º entran}} = \underbrace{Cp \times t \times 1/\bar{D}}_{\text{n.º salen}}$$

$Cp/(N-Cp)$ es la razón entre los casos y los “no casos”, equivalente a $P/(1-P)$. Esta medida se denomina *odds* de prevalencia. Por tanto:

$$I\bar{D} = \frac{P}{1-P} \quad [2-3]$$

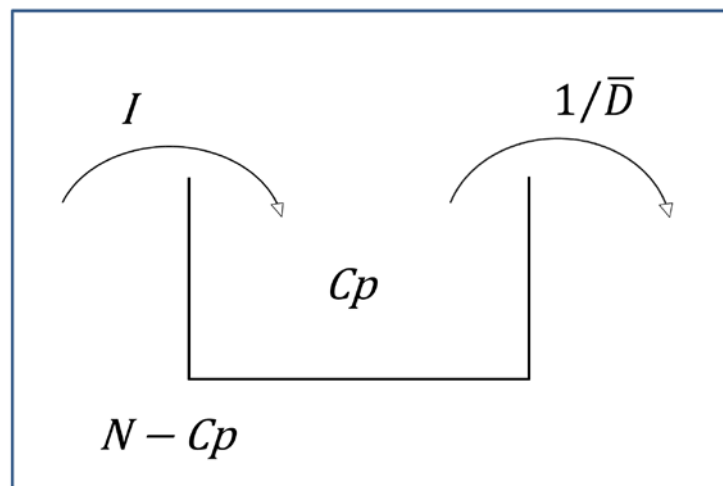
$$P = \frac{I\bar{D}}{1 + I\bar{D}}$$

Para prevalencias bajas ($<0,1$):

$$P \approx I\bar{D}$$

Los factores que afectan a la incidencia y a la duración de la enfermedad determinarán la prevalencia. Condiciones de corta duración, por ejemplo apendicitis, tendrán una prevalencia baja. Otras como diabetes pueden tener prevalencias altas aunque la incidencia no lo sea, debido a una gran duración de la enfermedad.

Figura 5. Prevalencia e incidencia. Representación de los flujos de enfermedad de una población estable.



Ejemplo 8. Tomando datos aproximados, la tasa de incidencia media de diabetes es de 9 casos por 1000 personas-año y su duración media de 12 años. La prevalencia es:

$$P = \frac{0,009 \text{ año}^{-1} \times 12 \text{ años}}{1 + (0,009 \text{ año}^{-1} \times 12 \text{ años})} = 0,121 \text{ (12,1\%)}$$

La fórmula aproximada infraestima algo la prevalencia:

$$P = 0,009 \text{ año}^{-1} \times 12 \text{ años} = 10,8\%$$

La prevalencia suele tener limitaciones para su utilización en estudios cuyo objetivo es comprobar hipótesis causales (estudios analíticos). Esto se debe a que uno de sus determinantes fundamentales es la duración de la enfermedad. Es posible por tanto que una diferencia de prevalencias entre dos poblaciones no pueda atribuirse a factores causales sino a una diferente duración media de la enfermedad. También se puede producir el fenómeno de la modificación de la exposición ocasionada por la propia enfermedad. Por ejemplo, si se quiere estudiar la relación entre ejercicio físico y enfermedad cardíaca, la utilización de casos prevalentes ocasionará que los enfermos/casos presenten valores muy reducidos de ejercicio físico debido a la enfermedad (causalidad inversa), modificando la verdadera asociación entre la exposición y la enfermedad.

1.6. Otras medidas

Odds. Son otra forma de expresar probabilidades. Son razones entre una proporción y su complementaria. Expresan cuánto más (o menos) probable es que ocurra un fenómeno (probabilidad p) comparado con que no ocurra (probabilidad $1-p$). Su rango está entre cero e infinito. Cuanto menor es el valor de la probabilidad más se acerca al valor de las odds correspondientes. La razón de odds (“odds ratio”) es muy utilizada en epidemiología como medida de asociación o efecto.

$$\text{odds} = \frac{p}{1-p} ; p = \frac{\text{odds}}{1 + \text{odds}}$$

Ejemplo 9. Si la probabilidad de tener al menos una caída en población que vive en residencias para mayores en un año es de 0,66, las odds de caídas en residencias son:

$$\text{odds} = \frac{0,66}{1 - 0,66} = 1,94$$

Es decir, la probabilidad de sufrir al menos una caída casi duplica la probabilidad de no sufrir ninguna.

Tasa de letalidad. Es la proporción de muertes entre los enfermos. Es la incidencia acumulada de muerte en personas con la enfermedad.

Tasa de ataque. Es la proporción de casos que se producen en una población cerrada y en un tiempo determinado, normalmente en contextos de brotes epidémicos. Es una incidencia acumulada con un tiempo de referencia variable determinado por la naturaleza de la enfermedad.

2. Medidas de efecto

Para poder estimar el efecto causal de una exposición en la ocurrencia de una enfermedad en poblaciones es necesario asumir que la enfermedad en personas expuestas puede ocurrir por otras causas diferentes de la exposición. Por tanto, si calculamos la tasa de incidencia de enfermedad en una población expuesta debemos conocer qué parte de esa tasa no se debe a la exposición. A la tasa de incidencia en población expuesta se debe restar *la tasa de incidencia que hubiera tenido esa misma población expuesta si no hubiera estado expuesta*. Medir esto no es posible. Lo que más se puede aproximar a este planteamiento teórico, llamado contra-fáctico (o contra-factual), ocurre en los *ensayos cruzados*, donde las mismas personas están expuestas unas veces y no expuestas otras (por ejemplo, a un tratamiento paliativo); pero incluso en estos casos las circunstancias no son las mismas, pues se comparan sucesos acaecidos en distintos periodos de tiempo. Como se verá en el capítulo de estudios experimentales, cuando se puede asumir que dos poblaciones son iguales en todo excepto en la exposición, las diferencias que se encuentren en sus tasas de incidencia se atribuirán necesariamente a la exposición.

2.1. Medidas absolutas. Diferencia de tasas y diferencia de riesgos

La diferencia entre la tasa de incidencia de una población expuesta I_1 y la tasa de incidencia que esperaríamos en esa misma población si no estuviera expuesta (I_1^c), $DT = I_1 - I_1^c$, representa una medida del *efecto absoluto* atribuible al factor de riesgo. Como se ha dicho I_1^c no es observable, pero es posible obtener una buena medida de su valor con una tasa de incidencia en una población *no expuesta* (I_0) que sea lo más parecida posible a la población expuesta. Si $I_1^c \neq I_0$ la asociación, medida como $DT = I_1 - I_0$, está confundida (por el efecto neto de factores de confusión, que son otras variables determinantes de la enfermedad cuya distribución es desigual entre expuestos y no expuestos) y por tanto no es causal. Aunque lo más habitual es que efectivamente $I_1^c \neq I_0$ existen varios métodos de control que permiten minimizar este problema. Por ejemplo, la estratificación, en la que se valora el efecto en subgrupos que sean homogéneos en factores de confusión con cierto peso.

Ejemplo 10. Deseamos conocer el efecto de la exposición a radiaciones en el desarrollo de leucemia. La tasa de incidencia de leucemia en una población expuesta a radiación (I_1) fue de 15,8 por 10^5 personas-año. La tasa equivalente en una población no expuesta (I_0) fue de 3,6 por 10^5 personas-año. El efecto absoluto atribuible a la radiación es:

$$DT = I_1 - I_0 = 0,000158 \text{ año}^{-1} - 0,000036 \text{ año}^{-1} = 0,000122 \text{ año}^{-1}$$

En una población de expuestos a radiación se producen 15,8 casos por cada 10^5 personas y año, de los que 12,2 casos por cada 10^5 personas y año se deben a radiación. Esto es válido si se puede asumir que la tasa en no expuestos estima bien la tasa que tendrían los expuestos a radiación si no estuvieran expuestos. En otras palabras: si las poblaciones de expuestos y no expuestos no difieren en otros determinantes de la enfermedad relevantes se puede considerar que la radiación es responsable de 12,2 casos por cada 10^5 personas y año.

Las consideraciones anteriores son equivalentes si las medidas de frecuencia comparadas son incidencias acumuladas. La medida del efecto absoluto en este caso es la *diferencia de riesgos*, $DR = IA_1 - IA_0$, que será una buena estimación del efecto causal si IA_0 es similar a IA_1^c (el riesgo que hubieran tenido los expuestos si no hubieran estado expuestos).

El *rango de valores* que puede tener una diferencia de tasas está entre $-\infty$ y $+\infty$; las unidades de medida son las mismas que tengan las tasas implicadas: año^{-1} , mes^{-1} , etc. Si se trata de una diferencia de riesgos, el rango está entre -1 y $+1$, sin unidades, ya que es una diferencia de proporciones. Si la diferencia es negativa significa que la exposición es preventiva. El valor que expresa la ausencia de efecto, el llamado *valor nulo*, es 0 para ambas medidas.

2.2. Medidas relativas. Razón de tasas. Razón de riesgos. Razón de odds

Las medidas de efecto absolutas son importantes en epidemiología y en salud pública porque expresan el impacto cuantitativo de la exposición, pero no expresan adecuadamente la *fuerza de la asociación*. Una diferencia de tasas puede ser grande en términos absolutos si las tasas implicadas son grandes, aunque sus diferencias relativas sean pequeñas, y a la inversa. Para tener una idea de la fuerza de la asociación hay que relacionar el efecto absoluto con una medida de la magnitud basal de la enfermedad (la incidencia en los no expuestos). Asumiendo que I_0 es similar a I_1^c , una medida del *efecto relativo* (ER) atribuible a la exposición es:

$$ER = \frac{I_1 - I_0}{I_0} = RT - 1$$

Es importante conocer cuánto más frecuente es la enfermedad en un grupo con respecto al otro. Normalmente se elimina de la ecuación la constante y se utiliza como medida relativa la *razón entre las dos tasas* (RT). Siguiendo con el *ejemplo 10*:

$$RT = \frac{I_1}{I_0} = \frac{0,000158 \text{ año}^{-1}}{0,000038 \text{ año}^{-1}} = 4,4$$

La tasa de incidencia en población expuesta a radiación es 4,4 veces la de población no expuesta. También es habitual utilizar el valor del ER , que es $4,4 - 1 = 3,4$, expresado en porcentaje: “La tasa de incidencia es un 340 % más alta en los expuestos con respecto a los no expuestos”. Esta medida del efecto relativo también se ha denominado *exceso de riesgo relativo* (entendiendo riesgo relativo como término genérico. En este ejemplo sería más correcto expresarlo como *exceso de razón de tasas*).

Si en lugar de tasas se utilizan incidencias acumuladas la medida de efecto es la *razón de riesgos*, con una interpretación equivalente:

$$RR = \frac{IA_1}{IA_0}$$

Un tercer tipo de medida relativa del efecto es la razón entre odds, llamada *odds ratio* (OR), que en los estudios de seguimiento es la razón entre las odds de enfermar de los expuestos y las de los no expuestos, también llamada *odds ratio de incidencia*:

$$OR = \frac{\frac{IA_1}{1 - IA_1}}{\frac{IA_0}{1 - IA_0}}$$

Cuanto menos frecuente sea la incidencia de enfermedad más se parecerán la *OR* y la *RR*. La *OR* es una medida menos intuitiva que la *RR* por lo que su uso en estudios de seguimiento es limitado.

Las medidas de efecto relativas, *RT* y *RR* y *OR*, carecen de unidades, sus posibles valores teóricos oscilan entre 0 e infinito (si bien el límite superior de la *RR* es $1/IA_0$) y el *valor nulo* es 1. Cuando la exposición incrementa el riesgo de adquirir la enfermedad se obtienen valores superiores a 1, y cuanto más altos, mayor es la fuerza de la asociación. Cuando la exposición disminuye el riesgo de adquirir la enfermedad (exposición preventiva) los valores son menores que 1 y cuanto más bajos, mayor es la fuerza de esta asociación preventiva.

Ejemplo 11. Estudio experimental para medir la efectividad de una intervención nutricional personalizada en pacientes hospitalizados en 6 centros. El grupo control recibió la alimentación habitual del hospital. En la tabla 4 se muestran los resultados en mortalidad por todas causas. El seguimiento fue de 30 días.

Tabla 4. Mortalidad a 30 días por todas las causas según intervención

	Intervención	Control
Casos	73	100
No casos	942	913
Total	1015	1013

$$IA_1 = \frac{73}{1015} = 0,072; IA_0 = \frac{100}{1013} = 0,099$$

$$DR = \frac{73}{1015} - \frac{100}{1013} = 0,072 - 0,099 = -0,027$$

$$RR = \frac{IA_1}{IA_0} = \frac{0,07}{0,1} = 0,73$$

$$OR = \frac{\frac{IA_1}{1-IA_1}}{\frac{IA_0}{1-IA_0}} = \frac{\frac{0,072}{1-0,072}}{\frac{0,099}{1-0,099}} = 0,71$$

La diferencia de riesgos se interpreta como que la intervención nutricional ha prevenido la muerte (a 30 días) en un 2,7 % de los pacientes del grupo de intervención.

El riesgo y las odds de muerte por cualquier causa en 30 días son un 27 % ($100 \times [1 - 0,73]$) y un 29 % ($100 \times [1 - 0,71]$) menor respectivamente en el grupo de intervención que en el control.

Si el tiempo de observación t es corto y la incidencia es baja, la razón de tasas y la razón de riesgos son casi iguales:

$$RR = \frac{IA_1}{IA_0} = \frac{I_1 t}{I_0 t} = \frac{I_1}{I_0} = RT$$

Por último, es importante tener en cuenta que, a diferencia de las medidas absolutas, las medidas relativas no informan sobre la magnitud del problema o impacto en la población. Su utilidad se limita a estudios de investigación causal. Una RR (o RT) alta puede ser responsable de muy pocos casos si la exposición en la población es rara. Y al revés. Además, a nivel individual tampoco informa del riesgo. Una RR alta no significa necesariamente un riesgo alto. Por ejemplo, una persona no expuesta puede multiplicar por 5 ($RR=5$) su probabilidad de enfermar con la exposición, pero si ésta era muy baja seguirá siendo baja, con lo que esto pueda implicar en la interpretación o en la toma de decisiones.

2.3. Medida del efecto en estudios de casos y controles

El diseño de casos y controles no permite que se realice una medición directa de la incidencia y por tanto no se pueden calcular medidas de efecto como diferencias o razones de tasas o riesgos. La estrategia en estos estudios consiste en comparar el nivel de exposición en los enfermos (casos) con el equivalente en un grupo de no enfermos (controles). Para ello se utiliza una razón de dos odds, llamada *odds ratio de exposición*. En un estudio de casos y controles se seleccionan $a + b$ casos de enfermedad y $c + d$ controles, personas sin la enfermedad, y se les mide la exposición:

Tabla 5. Notación de un estudio de casos y controles con una variable de exposición dicotómica

	Expuestos	No expuestos
Casos	<i>a</i>	<i>b</i>
Controles	<i>c</i>	<i>d</i>

Las *odds de exposición entre los casos*:

$$\frac{\text{probabilidad de estar expuesto}}{\text{probabilidad de no estar expuesto}} = \frac{a/a + b}{b/a + b}$$

y las *odds de exposición entre los controles*:

$$\frac{\text{probabilidad de estar expuesto}}{\text{probabilidad de no estar expuesto}} = \frac{c/c + d}{d/c + d}$$

Por tanto la *odds ratio*:

$$OR = \frac{\frac{a/a + b}{b/a + b}}{\frac{c/c + d}{d/c + d}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Esta *OR* se interpreta de forma similar a la *RT* o *RR*. Carece de unidades, el rango de valores está entre cero e infinito y el valor nulo que expresa no asociación es 1. Valores mayores de 1 indican una asociación positiva entre exposición y enfermedad y valores por debajo de 1 indican asociaciones preventivas. De hecho, el valor de la *OR* se utiliza como estimador de la *RT* o de la *RR* ya que, en general y dependiendo del diseño, es una buena aproximación de estos valores (3).

Ejemplo 12. Los resultados de un estudio que pretende evaluar la asociación entre antecedentes familiares de cáncer colorrectal y cáncer colorrectal son:

Tabla 6. Datos hipotéticos de un estudio de casos y controles sobre cáncer colorrectal y antecedentes familiares

	Antecedentes familiares	
	Sí	No
Casos	112	1472
Controles	108	2771

$$OR = \frac{112/1472}{108/2771} = \frac{112 \cdot 2771}{1472 \cdot 108} = 1,95$$

Las odds de exposición entre los casos casi duplica las de los controles. Este valor sirve como estimación de la RR y se interpreta de la misma forma: el riesgo de cáncer colorrectal de las personas con antecedentes familiares es un 95 % mayor que el de las personas que carecen de tales antecedentes.

2.4. Medida del efecto en estudios de prevalencia (transversales)

Si el diseño del estudio sólo permite estimar prevalencias, como ocurre en los estudios transversales, la *razón de prevalencias* (RP) puede utilizarse como estimación de la razón de tasas de incidencia bajo dos condiciones fundamentales: que la duración media de la enfermedad en los expuestos y no expuestos sea parecida y que las prevalencias sean bajas en ambos grupos.

A partir de la ecuación [2-3]:

$$P = I\bar{D}(1 - P)$$

La razón de prevalencias entre expuestos y no expuestos:

$$RP = \frac{I_1\bar{D}_1(1 - P_1)}{I_0\bar{D}_0(1 - P_0)} = RT \times \frac{\bar{D}_1}{\bar{D}_0} \times \frac{(1 - P_1)}{(1 - P_0)}$$

2.5. Asociación vs. efecto

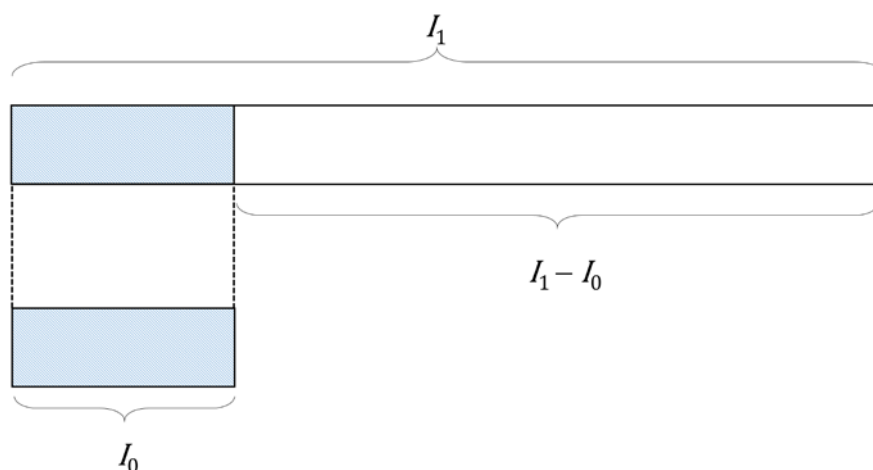
Una tendencia creciente entre investigadores y editores de revistas científicas es diferenciar los conceptos de asociación y efecto. Las medidas de diferencia o de razón se suelen interpretar como medidas de asociación o de efecto dependiendo del objetivo del estudio. Si el diseño permite atribuir causalidad con un grado razonable de confianza se denominan medidas de efecto. En caso contrario medidas de *asociación*. Por ejemplo, una *OR* se considera medida de asociación en un estudio descriptivo y medida de efecto si es un estudio analítico. Muchas veces la elección comporta un juicio de valor. En otras, la distinción es más clara. Se consideran claramente medidas de asociación las que comparan las frecuencias en poblaciones que no pueden tener escenarios contra-fácticos, como por ejemplo entre mujeres y hombres, o entre dos países. Otros autores plantean además que solo se puedan considerar medidas causales aquellas sobre exposiciones que puedan entenderse como eventuales intervenciones, razonablemente bien definidas —incluso aunque por motivos prácticos o éticos no pudieran realizarse (4)—. En los capítulos de estudios de intervención e inferencia causal en epidemiología se abordarán estas cuestiones en profundidad.

2.6. Proporciones atribuibles a la exposición

2.6.1. Proporción atribuible en expuestos

La diferencia de tasas de incidencia puede ser expresada como parte del total de la tasa de incidencia en expuestos (Figura 6). Tal medida se conoce como *proporción atribuible en expuestos* (PA_E).

Figura 6. Representación esquemática de la proporción atribuible en expuestos.



$$PA_E = \frac{I_1 - I_0}{I_1}$$

Ejemplo 13. Siguiendo con datos del ejemplo 10 sobre tasas de leucemia:

$$PA_E = \frac{0,000158 \text{ año}^{-1} - 0,000036 \text{ año}^{-1}}{0,000158 \text{ año}^{-1}} = 0,77$$

Es decir, el 77 % de la tasa de enfermedad en población expuesta es atribuible a la exposición.

La medida equivalente si se estiman riesgos es:

$$PA_E = \frac{DR}{IA_1} = \frac{IA_1 - IA_0}{IA_1}$$

Esta proporción expresa la *probabilidad* de que el factor de riesgo sea causa de la enfermedad en una persona expuesta¹. Esta probabilidad depende únicamente de la razón de riesgos:

$$PA_E = \frac{IA_1}{IA_1} - \frac{IA_0}{IA_1} = 1 - \frac{1}{RR} = \frac{RR - 1}{RR}; \quad (\text{si } RR \geq 1)$$

Conforme aumenta el valor de la *RR* la probabilidad se acerca a 1.

2.6.2. Proporción atribuible poblacional

Una medida similar a la PA_E se puede calcular para la población total, la proporción atribuible poblacional (PA_P):

$$PA_P = \frac{I_T - I_0}{I_T} \quad [2-4]$$

¹ Estrictamente esta *probabilidad de causación* se estima mediante la fracción etiológica, que es la fracción de casos expuestos para los cuales la exposición ha influido en el desarrollo de la enfermedad. Esta probabilidad no puede estimarse solo con datos epidemiológicos ya que depende además del modelo biológico de desarrollo de la enfermedad que se asuma. Si se utilizan riesgos la proporción atribuible en expuestos se puede considerar un límite inferior de la fracción etiológica. (Greenland S. *Concepts and pitfalls in measuring and interpreting attributable fractions, prevented fractions, and causation probabilities*. *Ann Epidemiol* 2015; 25: 155-161).

Donde I_T es la tasa de incidencia en la población total. La I_T puede expresarse como el promedio ponderado entre las tasas de incidencia en expuestos y no expuestos, donde los pesos son las proporciones de expuestos P_e y de no expuestos $1-P_e$:

$$I_T = P_e I_1 + (1 - P_e) I_0$$

Podemos sustituir en [2-4] y simplificar:

$$PA_p = \frac{RT - 1}{RT - 1 + \frac{1}{P_e}} ; \text{ (si } RT \geq 1 \text{)}$$

Ejemplo 14. Siguiendo con datos del ejemplo 10 de la leucemia, si los expuestos a radiación fueran el 2 % del total de la población:

$$PA_p = \frac{4,4 - 1}{4,4 - 1 + \frac{1}{0,02}} = 0,06$$

Es decir, en una población conjunta de expuestos y no expuestos el 6 % de todos los casos de leucemia son atribuibles a la radiación. En otros términos: si se actúa sobre la población eliminando totalmente la exposición se previene un 6 % de los casos.

Este parámetro tiene gran importancia en salud pública ya que nos ofrece una idea del número de casos atribuibles (potencialmente prevenibles) a una exposición. Como se aprecia, el valor de la PA_p también está en función de la RT , pero además depende de la cantidad de población expuesta (la *prevalencia de exposición*). Por muy fuerte que sea la asociación (RT muy alta) un factor de riesgo no será responsable de muchos casos en una población si hay muy pocas personas expuestas. Pero la exposición será responsable de muchos casos, aunque la RT sea baja, si está ampliamente presente en la población.

Ejemplo 15. Queremos evaluar el efecto de otro factor de riesgo, F , para leucemia (con datos del ejemplo 10). La fuerza de la asociación, con una RT de 1,5, es considerablemente más baja que la de la exposición a radiación. Sin embargo, F está presente en gran parte (40 %) de la población ($P_e = 0,4$) por lo que la proporción atribuible poblacional es:

$$PA_p = \frac{1,5 - 1}{1,5 - 1 + \frac{1}{0,4}} = 0,16$$

En contraste con el 6 % atribuible a radiación, F es responsable del 16 % de los casos.

Estas medidas son similares y con una interpretación equivalente si en lugar de tasas se miden riesgos. Además en los estudios de casos y controles se pueden obtener estimaciones de la proporción atribuible en expuestos y de la proporción atribuible poblacional mediante la OR , siempre y cuando la OR sea un buen estimador de la RR o RT :

$$PA_E = \frac{OR - 1}{OR} ; PA_P = \frac{OR - 1}{OR - 1 + \frac{1}{P_e}} ; \text{ si } OR \geq 1$$

Donde P_e es la proporción de expuestos en la población. Si no se dispone de este dato la proporción de expuestos entre los controles puede ofrecer una buena estimación si su selección fue razonablemente representativa (mejores controles comunitarios que hospitalarios).

Hay que recordar que el cálculo de estas fracciones atribuibles tiene sentido bajo la premisa de que las asociaciones son causales, y sobre factores modificables. Por otra parte, la fórmula para PA_P no es válida si se utilizan medidas de RT o RR u OR ajustadas por factores de confusión (5). Una fórmula que permite usar medidas relativas ajustadas es (6):

$$PA_P = P_c \frac{RME - 1}{RME} ; \text{ si } RME \geq 1$$

P_c es la proporción expuestos entre los enfermos y RME es la razón de mortalidad (o morbilidad) estandarizada. La RME es una razón de tasas ajustada mediante estandarización en que la distribución de potenciales factores de confusión utilizada como población estándar es la de la población expuesta (ver capítulo 6) y las tasas de referencia son las que tendría la población expuesta si no estuviera expuesta. Se puede sustituir la RME por cualquier medida de razón de riesgos o de tasas ajustadas que asuman homogeneidad de los efectos en los subgrupos.

Ejemplo 16. En un estudio para valorar la posible asociación entre incontinencia urinaria y mortalidad en personas que viven en residencias para mayores (7), se midió una razón de tasas de mortalidad ajustada por múltiples confusores de 1,24. La proporción de personas con incontinencia urinaria entre los casos de defunción fue del 57 %. La proporción de casos atribuible a incontinencia urinaria fue del 11 %.

$$PA_P = P_c \frac{RT_a - 1}{RT_a} = 0,57 \frac{1,24 - 1}{1,24} = 0,11$$

2.6.3. Proporción prevenida por la exposición

Si la exposición es preventiva ($I_1 < I_0$ o bien $IA_1 < IA_0$) se puede calcular la proporción prevenida en expuestos (PP_E), que es la proporción de casos prevenidos por la exposición del total de casos esperables si no se hubieran expuesto. Se calcula de forma similar a la PA_E . En el caso de un estudio que mida riesgos será:

$$PP_E = \frac{IA_0 - IA_1}{IA_0} = 1 - RR; \text{ (si } RR < 1)$$

Ejemplo 17. Un hipotético estudio de intervención trata de determinar el efecto que tiene una nueva vacuna de varicela. Finalizado el seguimiento los riesgos en expuestos y no expuestos a vacuna, así como la RR son:

$$IA_1 = \frac{9}{1200} = 0,0075; \quad IA_0 = \frac{130}{2000} = 0,065$$

$$RR = \frac{0,0075}{0,065} = 0,12$$

Es decir, los expuestos a la vacuna tienen un riesgo menor de enfermedad que los no expuestos. Si la población expuesta no hubiera estado expuesta se habría registrado un número adicional de casos hasta contabilizar una incidencia acumulada igual a la de la población no expuesta. El número de casos esperados en la población expuesta si no hubieran estado expuestos es:

$$\text{casos esperados en ausencia de exposición} = 0,065 \times 1200 = 78$$

La diferencia entre los 9 casos ocurridos y los esperados en ausencia de exposición es el n.º de casos que han sido prevenidos: $78 - 9 = 69$. La proporción prevenida es: $69/78 = 0,88$. Se obtiene el mismo resultado mediante la fórmula descrita anteriormente:

$$PP_E = 1 - RR = 1 - 0,12 = 0,88$$

También es habitual encontrar el resultado expresado en porcentaje: el riesgo en los expuestos es un 88 % menor que en los no expuestos. Cuando la exposición es una intervención la PP_E se denomina *eficacia de la intervención*, o *eficacia vacunal* si la intervención es una vacuna.

3. Referencias

1. Damián J, Pastor-Barriuso R, Valderrama-Gama E, de Pedro-Cuesta J. Factors associated with falls among older adults living in institutions. *BMC Geriatrics* 2013; 13:6.
2. Pollán M, Gustavsson P. High-risk occupations for breast cancer in the Swedish female working population. *Am J Public Health*. 1999; 89: 875-81.
3. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol*. 1982; 116(3): 547-53.
4. Hernán MA. Invited commentary: hypothetical interventions to define causal effects--afterthought or prerequisite? *Am J Epidemiol*. 2005; 162: 618-20.
5. Greenland S. Bias in methods for deriving standardized morbidity ratio and attributable fraction estimates. *Stat Med*. 1984; 3: 131-41.
6. Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol*. 1974; 99: 325-32.
7. Damián J, Pastor-Barriuso R, García López FJ, de Pedro-Cuesta J. Urinary incontinence and mortality among older adults residing in care homes. *J Adv Nurs*. 2017; 73: 688-99.

Capítulo 3.

Tipos de estudios epidemiológicos

JESÚS CASTILLA CATALÁN Y FERNANDO GARCÍA LÓPEZ

1. Diseño

2. Criterios de clasificación de los estudios epidemiológicos

- 2.1. Finalidad
- 2.2. Unidad de análisis
- 2.3. Direccionalidad
- 2.4. Selección de la muestra
- 2.5. Relación temporal o proximidad
- 2.6. Control de la asignación de los factores de estudio

3. Clasificación de los estudios epidemiológicos

- 3.1. Estudios descriptivos
- 3.2. Diseños observacionales incompletos o estudios de cribado de hipótesis
- 3.3. Estudios analíticos observacionales
- 3.4. Estudios experimentales o de intervención

4. Estudios epidemiológicos y grado de evidencia científica

5. Referencias

En un sentido amplio, el propósito de una investigación epidemiológica es conocer la distribución de un acontecimiento relacionado con la salud en una población, así como sus causas, los factores no causales que lo predicen y sus desenlaces. El diseño apropiado para una investigación epidemiológica determinada dependerá de los objetivos del estudio, de la información proporcionada por estudios previos, de los recursos disponibles y de otros factores.

El paradigma de la investigación epidemiológica es el estudio de intervención, que consiste en realizar observaciones bajo circunstancias controladas, en las cuales el investigador modifica las condiciones con el fin de establecer el efecto que tiene dicha modificación sobre las observaciones. De forma ideal, el estudio de intervención compara conjuntos de circunstancias que solo difieren en la presencia o ausencia del factor objeto de estudio. En la práctica, este estudio se considera aceptable si la variabilidad de otros factores distintos del que es objeto de estudio es suficientemente pequeña como para no afectar a los resultados de forma importante (ver capítulo de estudios de intervención).

1. Diseño

Cuando hablamos de la arquitectura o diseño de un estudio nos referimos a los procedimientos y métodos por los cuales se seleccionan los sujetos, se recoge y analiza la información y se interpretan los resultados. En otras palabras, el diseño es la estrategia que elegimos para cumplir los objetivos planteados.

Un estudio epidemiológico consta de varias fases: la definición del problema de investigación, la elección de la estrategia o diseño del estudio y la planificación de las actividades.

Toda investigación debe partir del conocimiento profundo y actualizado del tema que va a ser objeto de estudio. La primera fase consiste en definir el objeto de estudio, su justificación, el objetivo general y los objetivos específicos.

En la segunda fase se decide el diseño o la estrategia general que se va a emplear para dar respuesta a los objetivos planteados en la investigación. Esta decisión debe realizarse partiendo de un buen conocimiento de las características y ventajas de los distintos tipos de estudios, para primar aquellos que garanticen una mayor validez y adecuación al objetivo, pero teniendo también en cuenta los recursos humanos, económicos y de tiempo disponibles.

Una vez decidido el objetivo y el diseño se han de planificar las distintas actividades del estudio, que incluyen la definición de la población que se quiere estudiar, el método de selección de la muestra, los criterios de inclusión y exclusión, el periodo de estudio, la selección y definición de la variable de desenlace principal y de las demás variables, las fuentes de información que se van a utilizar, la determinación del tamaño muestral y el plan de procesamiento y análisis de los datos.

2. Criterios de clasificación de los estudios epidemiológicos

El interés de la clasificación de los estudios epidemiológicos es indudable. Proporciona una terminología concisa, necesaria para una comunicación eficaz entre los profesionales. A la vez, orienta sobre las limitaciones que pueden aparecer en función del tipo de estudio y a identificar el diseño más adecuado para cada objetivo.

Existen muchas clasificaciones de los estudios epidemiológicos, pero la mayoría se basan en algunos criterios básicos comunes: la finalidad del estudio, la unidad de análisis, la direccionalidad, la forma de selección de la muestra, la relación temporal y el control de la asignación del factor a estudio por parte del investigador. A continuación, se describen de forma sucinta cada uno de estos criterios.

2.1. Finalidad

Los estudios epidemiológicos se clasifican según su finalidad en descriptivos y analíticos. Los *estudios descriptivos* son los que estudian la frecuencia y distribución de los fenómenos de salud y enfermedad, mientras que los *analíticos* se dirigen a evaluar presuntas relaciones de causa-efecto. En otras palabras, los estudios descriptivos tratan de dar respuesta a preguntas sobre el ¿dónde?, ¿cuándo?, ¿quiénes?, ¿cuántos? y ¿cómo?; mientras que los estudios analíticos tienen por objeto responder al ¿por qué? de los fenómenos de salud y enfermedad (Figura 3.1).

Entre la epidemiología descriptiva y la analítica existen múltiples conexiones. Aunque la observación y descripción de la realidad en muchos casos es un fin en sí mismo, también puede ser el primer paso para el estudio de hipótesis causales. En los estudios analíticos es frecuente que haya un componente descriptivo inicial. Los estudios descriptivos también pueden analizar presuntas relaciones de causa-efecto, pero, debido a sus limitaciones en el diseño, sus resultados tienen un carácter meramente exploratorio.

2.2. Unidad de análisis

Todos los estudios epidemiológicos tienen en común analizar problemas de salud en la población; no obstante, la unidad de análisis puede ser:

- *el individuo*, cuando se utilizan datos desagregados de cada sujeto, o
- *las poblaciones*, cuando se utilizan datos agregados de grupos de individuos (países, regiones, ciudades, distritos, familias, colegios, empresas, etc.). Estos suelen llamarse *estudios ecológicos*.

2.3. Direccionalidad

Kramer y Boivin (1) proponen clasificar los estudios según el orden en que se investiga la asociación entre la causa y el efecto. Consideran tres posibilidades:

- Hacia delante, o *secuencia desde la causa al efecto*. Se seleccionan los sujetos en función de la exposición, y tanto en las personas expuestas como en las no expuestas se evalúa la aparición del efecto. Esta es la característica principal de los *estudios de cohortes*.
- Hacia atrás, o *secuencia desde el efecto a la causa*. Se seleccionan personas, unas que presentan el efecto y otras no, y se evalúa la presencia de la exposición en todos. Un ejemplo de ello serían los *estudios de casos y controles*.
- Simultánea o *sin direccionalidad*. La exposición y el efecto son evaluados a la vez. Estos serían los *estudios transversales*.

Las dos primeras situaciones, en las que sí hay direccionalidad, se engloban bajo el término de *estudios longitudinales* y permiten establecer el orden en el que se producen los acontecimientos en el tiempo, que es un argumento indispensable para demostrar relaciones de tipo causal. No obstante, este argumento solo puede considerarse cumplido cuando el tiempo entre la exposición y el efecto abarca el periodo de inducción correspondiente.

2.4. Selección de la muestra

Según el criterio utilizado para elegir a partir de la población diana a los sujetos que conformarán la muestra a estudiar, existen dos posibilidades:

- Muestreo probabilístico. Consiste en seleccionar una muestra con la intención de que sea representativa de la población diana. Es propio de los estudios descriptivos de prevalencia.
- Muestreo no probabilístico, o de conveniencia. Cuando el efecto que se quiere estudiar es poco frecuente puede ser más eficiente muestrear per-

sonas que presentan el efecto y, después, buscar el grupo de comparación más adecuado. Por el contrario, cuando la exposición a estudiar es rara, puede interesar buscar directamente personas expuestas y compararlas con un grupo de personas no expuestas. Esta última situación se produce en el estudio de exposiciones que ocurren en determinadas actividades laborales. Las muestras de conveniencia son más apropiadas para los estudios analíticos, que analizan hipótesis causales.

2.5. Relación temporal o proximidad

El tiempo transcurrido desde que se produjeron los hechos que se analizan hasta el momento en el que se realiza el estudio puede influir en la validez de la información. En función de esta relación temporal los diseños se clasifican en:

- *Históricos o retrospectivos.* Estudian hechos ocurridos antes del comienzo del estudio. La información puede obtenerse a partir de registros existentes, como por ejemplo en el caso de los estudios de cohortes retrospectivas, o indagando en las entrevistas sobre hechos ocurridos en el pasado.
- *Concurrentes o prospectivos.* Consideran únicamente eventos que se producen a partir del momento de inicio del estudio.
- *Mixtos.* Estudian tanto hechos históricos como concurrentes.

Conviene no confundir la relación temporal con la direccionalidad. Así, por ejemplo, un estudio de casos y controles puede ser prospectivo, si se realiza sobre casos incidentes, o retrospectivo, si incluye casos prevalentes.

2.6. Control de la asignación de los factores de estudio

Los *estudios experimentales o de intervención* son aquellos en los que el investigador controla la asignación de la exposición a estudio. Todos ellos son de tipo analítico y, por razones éticas, suelen limitarse exclusivamente a evaluar intervenciones terapéuticas o preventivas.

Los *estudios no experimentales*, comúnmente llamados *observacionales*, son todos aquellos en los que el investigador no controla la asignación de la exposición, sino que se limita a analizar factores cuya presencia o ausencia en los individuos se ha producido por un motivo independiente a la investigación.

3. Clasificación de los estudios epidemiológicos

Existen muchas clasificaciones de los estudios epidemiológicos, la mayoría de las cuales se basan en los criterios mencionados en el apartado anterior. En las tablas 3.1., 3.2., 3.3. y 3.4., y en las figuras 3.1. y 3.2. se muestran varios ejemplos. Las principales diferencias entre las diversas clasificaciones radican en la importancia y prioridad que dan a cada criterio y en discrepancias en la terminología utilizada. Esto pone de manifiesto la dificultad para establecer un esquema que permita clasificar la enorme diversidad de estudios posibles. Por otra parte, no es raro que en una misma investigación haya más de un objetivo y coexistan varios diseños diferentes. Además, existen diseños mixtos, híbridos y situaciones intermedias que complican la aparente simplicidad de las clasificaciones.

Figura 1. Clasificación de los estudios epidemiológicos según las preguntas de investigación. Esquema adaptado a partir de materiales docentes de J. M. Martín Moreno.

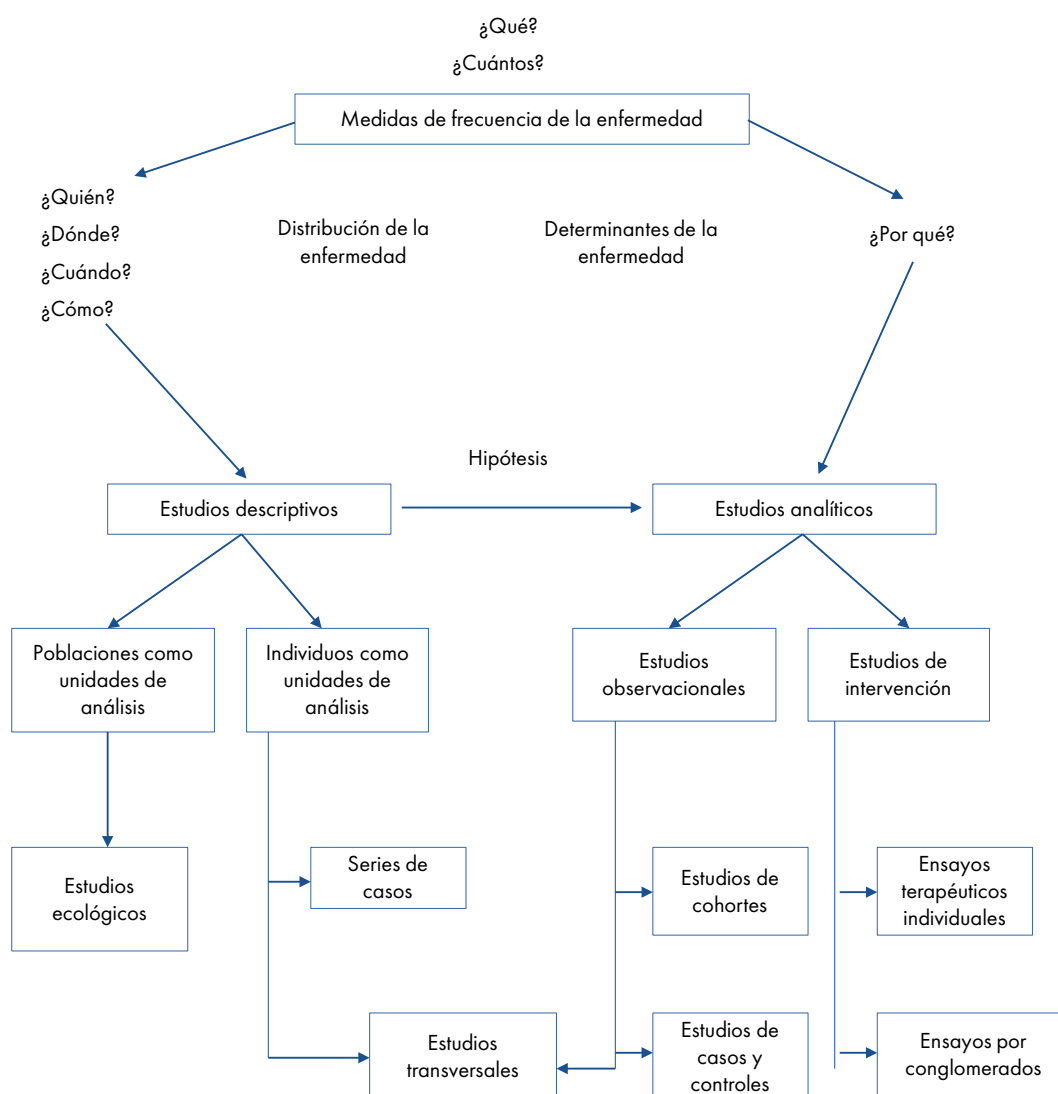
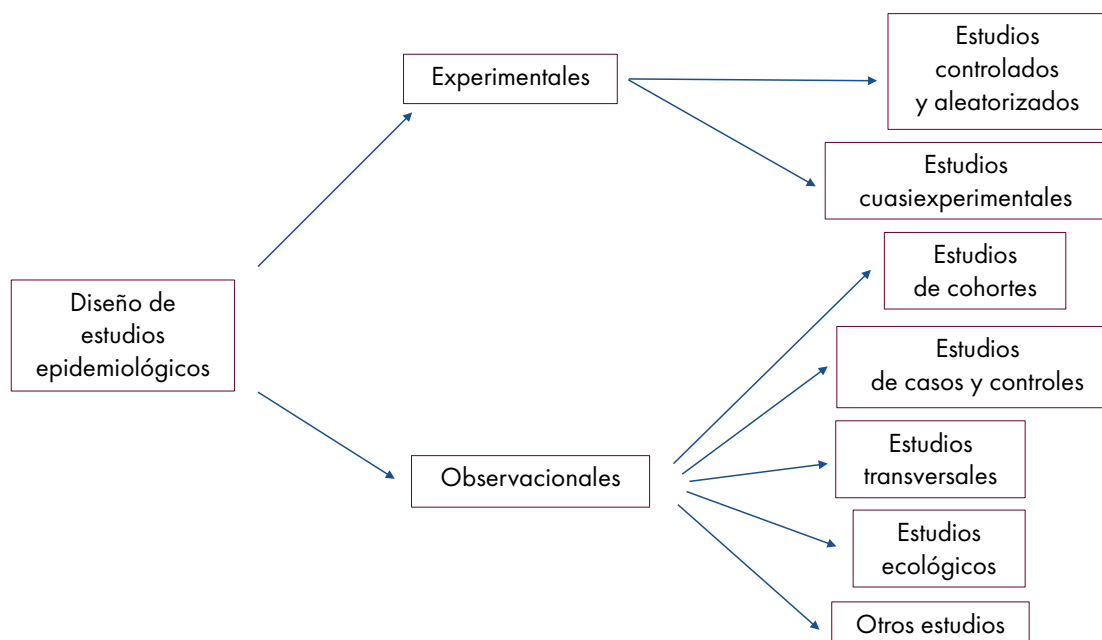


Figura 2. Clasificación de los estudios epidemiológicos según el control de la intervención por el investigador.**Tabla 1. Clasificación de los estudios epidemiológicos**

Estudios descriptivos	Estudios observacionales de cribado de hipótesis. Diseños incompletos	Estudios analíticos observacionales
- Estudio de una serie de casos - Transversal - Longitudinal - Estudios descriptivos de datos agregados - Análisis geográficos - Análisis de series temporales - Análisis en función de otras variables	- Estudios ecológicos de correlación - Estudios de mortalidad (o morbilidad) proporcional	- Estudios transversales - Estudios de casos y controles - Incidentes - Prevalentes - Estudios de cohortes - Diseños híbridos - Estudios de cohorte y casos o <i>case-cohort</i> (muestreo de controles a partir de la <i>cohorte fuente</i>) - Estudios de casos y controles anidados en una cohorte (muestreo de densidad de incidencia de los controles) - Estudios experimentales - Ensayos controlados y aleatorizados * Ensayos clínicos * Ensayos preventivos - Ensayos de campo - Ensayos comunitarios - Estudios cuasiexperimentales o de intervención no aleatorizados

Tabla 2. Clasificación de los estudios observacionales (modificada a partir de Kleinbaum DG et al., 1982) (2)

Diseños básicos	Diseños incompletos	Diseños híbridos
- Estudios de cohortes - Estudios de casos y controles - Estudios transversales	- Estudios ecológicos - Estudios de mortalidad o morbilidad proporcional - Estudios de agregación temporoespacial	- Estudios de cohorte y casos (muestreo de controles a partir de la cohorte al inicio del seguimiento) - Estudios de casos y controles anidados en una cohorte (muestreo de densidad de incidencia de los controles)

Tabla 3. Cuatro tipos de estudios epidemiológicos básicos (modificada a partir de Pearce N, 2012) (3)

Desenlace del estudio	Muestreo en el desenlace	
	No	Sí
Incidencia	Estudios de incidencia	Estudios de casos y controles de incidencia
Prevalencia	Estudios de prevalencia	Estudios de casos y controles de prevalencia

Tabla 4. Clasificación de los estudios epidemiológicos en función del control de la exposición, el objetivo y la unidad de análisis

Según el control de la exposición	Objetivo	Unidad de análisis	
		Individuo	Población
Experimentales	Analítico	Ensayos clínicos Ensayos de campo	Ensayos por conglomerados
Observacionales	Analítico (comprobación de hipótesis)	Estudios de cohortes Estudios de casos y controles Estudios transversales analíticos	
	Diseños incompletos (cribado de hipótesis)	Estudios de mortalidad o morbilidad proporcional	Estudios ecológicos de correlación
	Descriptivo	Encuestas de prevalencia descriptivas Series de casos	Estudios descriptivos con datos agregados

3.1. Estudios descriptivos

Los estudios descriptivos son los que tienen por objeto observar y describir los acontecimientos relacionados con la salud en función de las variables de tiempo, lugar y persona. Son, con diferencia, los estudios más frecuentes y la aplicación más habitual de la epidemiología.

La gran mayoría de estos estudios se realizan con la finalidad de aportar información sobre aspectos concretos de la realidad, que sirva de apoyo a la toma de decisiones en los distintos ámbitos relacionados con la salud: la actividad asistencial, la planificación sanitaria, la gestión de recursos de salud, las intervenciones en salud pública y las políticas sanitarias. En la medida que los estudios descriptivos suponen la observación y descripción de fenómenos, ofrecen también una fuente de hipótesis para la investigación epidemiológica.

Los estudios descriptivos pueden utilizar información existente (fuentes secundarias) o recoger nueva información a tal efecto. Entre las posibles fuentes de datos están las historias clínicas, cuestionarios específicos, resultados de analíticas de laboratorio, sistemas de vigilancia, estadísticas vitales, bases de datos de actividad asistencial, registros de morbilidad, datos de consumos de recursos sanitarios, etc.

Algunos estudios descriptivos tienen como unidad de análisis el individuo, entre los que están *la descripción de un caso*, *la serie de casos* y los *estudios transversales*. Otros tienen como unidad de análisis la población o el grupo de individuos y son los llamados *estudios ecológicos*.

3.1.1. Descripción de un caso o de una serie de casos

Estos diseños describen la experiencia de un único paciente o de un grupo de pacientes con diagnóstico similar, que presentan alguna característica o curso clínico inusual. Pueden ser el punto de partida para una nueva hipótesis. Estos estudios pueden generarse a partir de la actividad asistencial o de la notificación a los sistemas de vigilancia epidemiológica. El paso fundamental en ellos es establecer la definición de caso y los criterios de inclusión y exclusión. Su principal limitación es la falta de un grupo control, lo que impide evaluar hipótesis causales.

Estos estudios suponen una tercera parte de las publicaciones biomédicas. En muchos casos quedan en la bibliografía como la descripción de una rareza, pero en algunas ocasiones han sido el primer paso para la identificación de una enfermedad emergente o para un nuevo avance científico. La descripción de una serie de casos puede ser la primera constancia de un brote epidémico o de una nueva epidemia. Un ejemplo de ello fue la descripción en 1981 en Los Ángeles de 5 casos de neumonía por *Pneumocystis carinii* en hombres homosexuales previamente sanos. Esta serie de casos fue la primera descripción del sida y el primer paso para

el descubrimiento del virus de la inmunodeficiencia humana. Otro ejemplo fue la descripción en España de un aumento de casos de neumonía atípica que no respondían al tratamiento y que más tarde permitió la identificación del síndrome por consumo del aceite tóxico.

La descripción de un caso o de una serie de casos puede plantearse de forma transversal o longitudinal. Las series longitudinales se componen de la cohorte de sujetos en riesgo con el fin de cuantificar la incidencia de una enfermedad, o la cohorte de sujetos enfermos para describir su historia natural.

3.1.2. Estudios descriptivos con datos agregados

También son llamados *estudios ecológicos descriptivos*. Describen la frecuencia de un determinado acontecimiento relacionado con la salud en la población, su tendencia en el tiempo, su distribución en el espacio o en función de otras características de las personas como el sexo, la edad, la profesión, el nivel educativo, la clase social, etc. Son abordajes muy frecuentes en el campo de la salud pública, como apoyo a actividades de planificación y gestión sanitarias.

Las posibilidades de estos estudios están condicionadas a la existencia de sistemas de información y a la calidad de sus datos. Las fuentes de información pueden ser las estadísticas de mortalidad, los registros de morbilidad, los sistemas de vigilancia epidemiológica u otros sistemas de información sanitaria. Permiten obtener medidas de frecuencia absolutas y relativas, de incidencia, prevalencia o mortalidad, y pueden establecer comparaciones entre grupos. Los cambios en la frecuencia del acontecimiento estudiado a lo largo del tiempo, las diferencias geográficas, o las diferencias en función de las características de las personas, son fuentes de hipótesis para estudios analíticos posteriores.

El planteamiento general de estos estudios es sencillo y requiere pocos medios y poco tiempo para su ejecución. Son cada vez más frecuentes, al disponerse de mejores sistemas de información sanitaria y de nuevas herramientas metodológicas. Los estudios descriptivos en función de la variable tiempo han encontrado un amplio desarrollo a través de los *análisis de series temporales* y de *edad-periodo-cohorte*. Los estudios en función de la variable espacio han expandido sus posibilidades gracias a los *sistemas de información geográfica*.

En este apartado también se incluyen los estudios que exploran posibles agregaciones temporales o espaciales de casos. Cuando se detecta una agregación de casos de una enfermedad en el tiempo o en el espacio, por encima de lo esperable por el azar, es muy sugerente de un foco etiológico común, probablemente infeccioso o ambiental. La detección de una agregación de casos puede aportar hipótesis para la investigación etiológica y puede ayudar a identificar el foco común para actuar sobre él y controlar la aparición de nuevos casos.

3.1.3. Estudios descriptivos transversales, o de prevalencia

Los *estudios transversales* o *de prevalencia* (aunque algunos autores usan esta denominación solo cuando la variable de interés es dicotómica) pueden realizarse con una finalidad analítica, pero en muchos casos se limitan a objetivos meramente descriptivos. Por ello, dependiendo de cada caso, podrían ser clasificados como estudios descriptivos o analíticos, e incluso algunos autores han llegado a incluirlos en una categoría intermedia de estudios mixtos. Su característica principal es que no tienen direccionalidad, es decir, toda la información recogida hace referencia a un mismo momento en el tiempo. Con frecuencia la selección de la muestra se hace de forma representativa de la población diana, lo que facilita el que los resultados sean generalizables a dicha población. Ofrecen una visión instantánea de lo que ocurre en la población en un momento dado. La medida de frecuencia que proporcionan es la prevalencia.

A este tipo de estudios pertenecen las encuestas que se realizan como apoyo a las actividades de planificación sanitaria de diversa índole, como las encuestas de salud, las encuestas de satisfacción de los usuarios, etc. En ocasiones, las variables de estudio son resultados de determinaciones de laboratorio, como es el caso de las *encuestas seroepidemiológicas*, cuyo objeto es conocer la prevalencia de marcadores serológicos de distintas infecciones en la población diana.

Entre los estudios descriptivos transversales se incluyen los que evalúan métodos diagnósticos, que comparan los resultados de una técnica diagnóstica nueva con los de la técnica de referencia y, también, los estudios que evalúan la concordancia entre técnicas o entre observadores.

3.2. Diseños observacionales incompletos o estudios de cribado de hipótesis

En este apartado nos referimos a algunos estudios epidemiológicos que se dirigen a explorar presuntas relaciones de causa-efecto, si bien, debido a las limitaciones de su diseño, sus resultados se consideran exploratorios. Entre estos podrían incluirse los estudios ecológicos y los de morbilidad o mortalidad proporcional. Muchos manuales de epidemiología incluyen este grupo de estudios en el apartado de los descriptivos. Sin embargo, Rothman *et al.* (4) se refieren a estos como estudios de cribado de hipótesis, ya que cumplirían la función de seleccionar hipótesis de trabajo para posteriores estudios analíticos. Kleinbaum *et al.* (2) utilizan el término de diseños observacionales incompletos.

3.2.1. Estudios ecológicos de correlación

Los estudios ecológicos se caracterizan porque las unidades de análisis son grupos de individuos o poblaciones. Utilizan una medida de la frecuencia de la enfermedad en cada una de las poblaciones de estudio, como por ejemplo la prevalencia, y evalúan su asociación con algún factor de interés como la edad media, la fecha de estudio, la tasa de utilización de servicios de salud, o el consumo promedio de determinados alimentos, medicamentos u otros productos. La medida de asociación que se obtiene es el coeficiente de correlación y por ello también se les llama estudios de correlación.

Los grupos de individuos que se utilizan como unidad de estudio son habitualmente los habitantes de distintos países, regiones o ciudades, pero también pueden compararse cohortes de nacimiento o grupos más pequeños como colegios, empresas o familias. Cuando se compara la población de un mismo lugar en distintos momentos del tiempo se habla de *estudios de ecológicos de tendencia temporal*.

Un ejemplo de estudio ecológico es el análisis de la correlación entre el consumo per cápita de cigarrillos en diversos países y sus correspondientes tasas de mortalidad por enfermedad coronaria.

La principal ventaja de estos diseños es que se realizan sobre fuentes de información secundarias, ya disponibles, lo que hace que sean rápidos de ejecución y económicos. Su principal limitación es la imposibilidad de establecer la relación causal entre la exposición y el efecto a nivel de los individuos. Puede ocurrir que se observe una asociación cuando se analizan poblaciones, sin que exista realmente en el nivel de los individuos, situación que se denomina *falacia ecológica*.

3.2.2. Estudios de mortalidad (o morbilidad) proporcional

Son estudios que tratan de evaluar la asociación entre un factor y un acontecimiento, que puede ser la morbilidad o la mortalidad por una determinada causa.

Se realizan a partir de información de registros de mortalidad, o de registros de morbilidad como los de cáncer, de sida, de enfermedades de declaración obligatoria, de altas hospitalarias, etc. En un estudio de mortalidad proporcional la comparación se establece entre la proporción de las muertes en sujetos expuestos que es debida a una determinada causa, con la misma proporción en los sujetos no expuestos. Para algunos autores los estudios proporcionales deben plantearse como estudios de casos y controles (Rothman *et al.*) (4), donde los controles son enfermos o fallecidos por otras causas.

La aplicabilidad de estos estudios está limitada en función de las variables que se recogen en los registros. La medida de efecto que se obtiene es la razón de mortalidad (o morbilidad) proporcional. Esta medida es un buen estimador de la ra-

zón de tasas cuando se cumple el supuesto de que la morbilidad o mortalidad por las restantes causas es independiente de la exposición; no obstante, este supuesto es difícil de demostrar. Otra limitación de los estudios de mortalidad proporcional es que no permiten diferenciar si la exposición se asocia a la aparición de la enfermedad o a su evolución letal en los que la padecen.

3.3. Estudios analíticos observacionales

Los estudios analíticos son aquellos cuyo objetivo es evaluar hipótesis y responder al porqué de los fenómenos de salud y enfermedad. En función de la actitud que adopta el investigador, los estudios analíticos se clasifican en dos grandes grupos: los estudios observacionales y los experimentales o de intervención. En los observacionales, el investigador no interviene en la exposición de los sujetos de estudio. Se limita a medir dicha exposición y estimar asociaciones con enfermedades (o desenlaces).

La mayor parte de los autores distinguen tres tipos de *diseños observacionales básicos*: los diseños transversales, los de casos y controles, y los de cohortes. Además, hay diseños híbridos que combinan elementos de distintos diseños básicos. En la figura 3.3. se muestran de forma esquemática un ejemplo de los tres diseños básicos y en la tabla 3.5. un análisis comparativo de las ventajas e inconvenientes de los principales diseños analíticos.

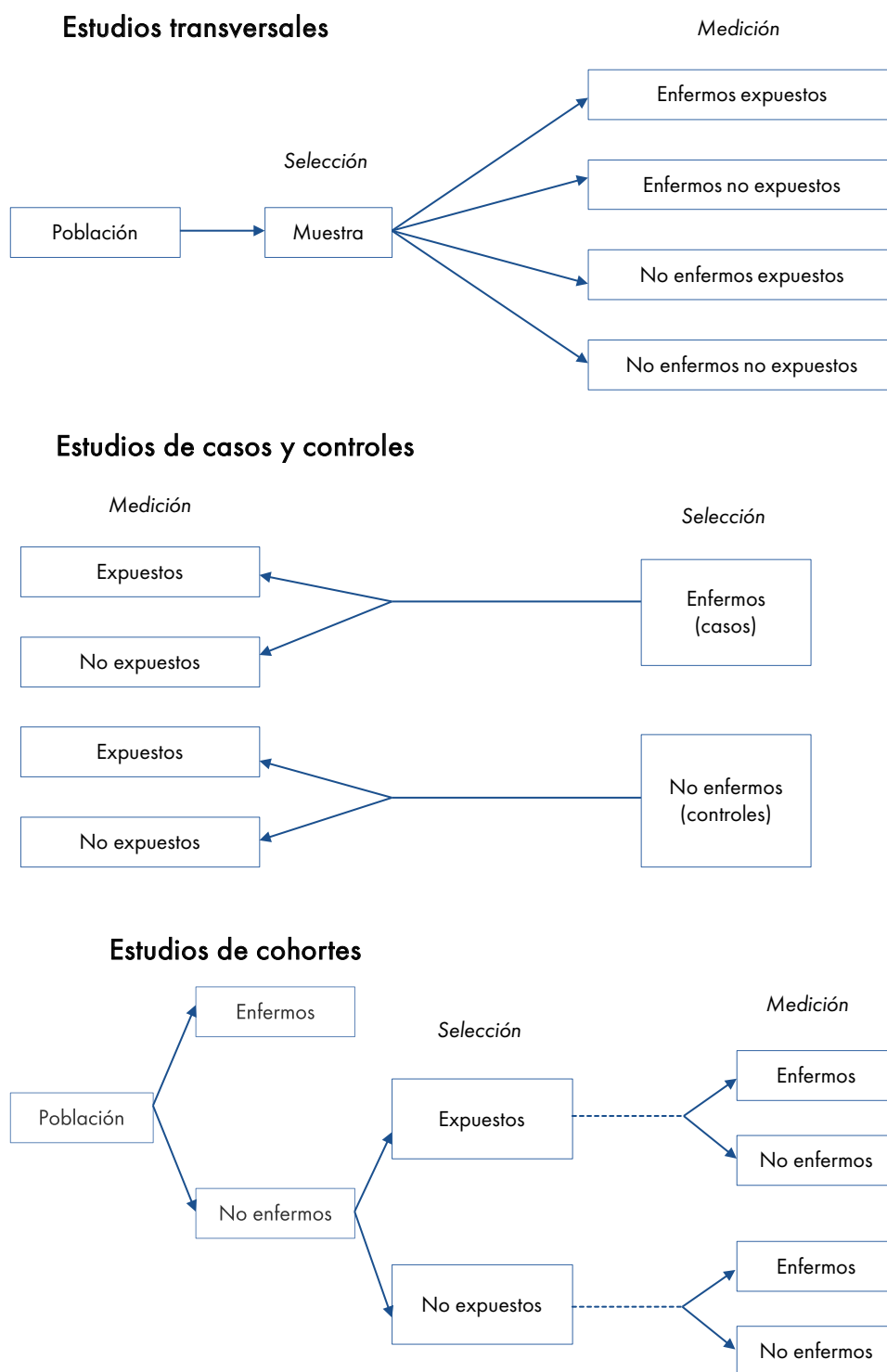
Figura 3. Esquema de un ejemplo de diseños transversal, de casos y controles y de cohortes.

Tabla 3.5. Ventajas e inconvenientes de distintos diseños epidemiológicos analíticos

	Estudios transversales	Estudios de casos y controles	Estudios de cohortes	Estudios experimentales
Ventajas	• Especialmente indicados en enfermedades crónicas • Resultados fácilmente generalizables a la población • Rápidos de realizar y baratos • Aportan medida de la prevalencia, útil para tareas de administración sanitaria • Permiten analizar varios efectos simultáneamente	• Rápidos de realizar y baratos • Adecuados para enfermedades con largos periodos de latencia • Óptimos para enfermedades raras • Permiten analizar varios factores etiológicos de la enfermedad	• Óptimos para exposiciones raras • Permiten estudiar varios efectos de una misma exposición • Establecen la relación temporal entre la exposición y el efecto • Permiten medir incidencia • En las cohortes concurrentes se minimizan los sesgos en la medida de la exposición	• Permiten establecer la secuencia temporal entre la exposición y el efecto • Permiten un mayor control de la exposición, de la confusión y de algunos sesgos
Inconvenientes	• No permiten establecer la relación temporal entre la causa y el efecto • No permiten distinguir entre factores de riesgo y factores pronósticos • Posibles sesgos en la medida de la exposición • No útiles en enfermedades raras	• Ineficientes para exposiciones raras • No permiten obtener tasas de incidencia • Puede ser difícil establecer la relación temporal de los sucesos • Tendencia a sesgos de selección y del recuerdo	• Ineficientes para enfermedades raras • Los estudios prospectivos son costosos y de larga duración • Las cohortes históricas requieren la existencia de registros disponibles • La validez puede afectarse por pérdidas de seguimiento	• Limitados a evaluar medidas preventivas o terapéuticas, por razones éticas • Menor validez externa por realizarse en poblaciones muy seleccionadas • Solo estudian una intervención • Elevado coste

3.3.1. Estudios transversales

Los estudios transversales o de prevalencia se utilizan en ocasiones para evaluar hipótesis causales. La selección de la muestra suele ser representativa de la población de estudio, lo que facilita que los resultados sean generalizables al conjunto.

Se caracterizan por no tener direccionalidad: tanto la evaluación de la exposición como la del efecto hacen referencia a un mismo momento en el tiempo (Figura 3.2.). Esto impide establecer la relación temporal entre ambos, lo que limita la aportación de este diseño a la investigación de relaciones causales

Otros posibles problemas son que ante una asociación puede no ser fácil distinguir entre un factor de riesgo y un factor pronóstico, y que los casos con mayor duración tienden a estar sobrerrepresentados. Por todos estos motivos se atribuye a los estudios transversales menor valor que a los otros diseños analíticos a la hora de evaluar asociaciones causales.

No obstante, los estudios transversales pueden ser adecuados para evaluar exposiciones que no cambian en el tiempo, como el grupo sanguíneo, o cuando la exposición actual es un buen indicador de la exposición relevante respecto al posible efecto, lo cual es más probable si la exposición es constante en el tiempo. En algunos temas, la investigación etiológica suele hacerse sobre casos prevalentes por la dificultad de establecer el comienzo de los procesos, como ocurre con las enfermedades congénitas, la hipertensión arterial, la diabetes, etc.

3.3.2. Estudios de casos y controles

En los estudios de casos y controles se seleccionan dos grupos de sujetos, unos con el efecto que se estudia (casos) y otros sin él (controles). La direccionalidad del estudio se establece hacia atrás, ya que en un segundo paso se indaga en cuáles de los sujetos estuvo presente la exposición que se estudia (Figura 3.3).

En función de la relación temporal puede ser:

- *retrospectivo o histórico*, si se incluyen casos prevalentes,
- *prospectivo o concurrente*, si se incluyen solo casos incidentes, y
- *mixto*, si se incluyen ambos tipos de casos.

Los casos prevalentes están más expuestos a posibles sesgos de supervivencia selectiva y de memoria. En los estudios de casos y controles no es posible obtener directamente medidas de incidencia y se recurre a la *odds ratio* como medida de asociación que se aproxima a la razón de tasas, o de riesgos.

Los estudios de casos y controles son especialmente eficientes para el estudio de enfermedades poco frecuentes o con largos periodos de latencia.

3.3.3. Estudios de cohortes

La característica que los distingue es que la selección de sujetos de estudio se realiza en función de la exposición en estudio. Los acontecimientos se estudian en la misma secuencia en la que se produjeron, lo que hace que este diseño sea más

intuitivo y fácil de entender. La selección de la muestra puede hacerse de forma representativa de la población diana o de conveniencia, en función de criterios pragmáticos, como por ejemplo la exposición en las cohortes ocupacionales. Esta última posibilidad es especialmente útil cuando se estudian exposiciones raras.

En función de la relación temporal entre los hechos que se estudian y el momento de realización de la investigación, los estudios de cohortes pueden ser históricos, concurrentes o mixtos.

- Los *estudios de cohortes históricas*, también son llamadas *retrospectivas*, se basan en información recogida en registros.
- Los *estudios de cohortes concurrentes o prospectivos* solo tienen en cuenta los acontecimientos que se producen a partir del comienzo del estudio.
- Los *estudios de cohortes mixtas* son los que combinan información prospectiva y retrospectiva.

Inicialmente todos los sujetos están libres del efecto y se diferencian según la exposición. A partir de ese momento se requiere contemplar el seguimiento de todos los sujetos durante un intervalo de tiempo suficiente para dar margen al periodo de inducción del efecto que se analiza. Al final del seguimiento se evalúa la aparición del efecto en los dos grupos (Figura 3.3.).

Los estudios de cohortes permiten establecer la relación temporal entre la exposición y el efecto y obtener medidas de incidencia. Además, si las cohortes son concurrentes, se minimizan los sesgos en la medida de la exposición.

3.3.4. Diseños híbridos

Bajo este epígrafe se incluyen los *estudios de cohorte y casos* (*case-cohort* en inglés) y los *estudios de casos y controles anidados en una cohorte*. En ambos se parte de un estudio de cohortes. En el análisis se incluyen todos los casos y, como grupo de comparación, se toma una muestra de la cohorte. De esta forma se reduce considerablemente el número de sujetos que es necesario estudiar, sin que ello suponga una disminución importante de la potencia estadísticas del estudio. Estos diseños consiguen mejorar mucho la eficiencia cuando la medición de la exposición o de otras variables de análisis es costosa.

En estos diseños híbridos la medida de asociación es la *odds ratio*, que estima la razón de riesgos en los *estudios de cohorte y casos* y la razón de tasas de en los *estudios de casos y controles anidados en una cohorte*.

3.4. Estudios experimentales o de intervención

Los estudios experimentales o de intervención son aquellos en los que la exposición es asignada por el investigador en función de lo establecido previamente en el protocolo. Cuando la asignación se realiza en función de otros criterios, como por ejemplo la aplicación de un tratamiento siguiendo criterios terapéuticos, estaremos ante un estudio observacional, de los también denominados pseudoexperimentales.

3.4.1. Ensayos controlados y aleatorizados

Con el fin de garantizar que los grupos sean lo más comparables posible y así minimizar la confusión, lo ideal es que la asignación de la exposición se efectúe de forma aleatoria, aunque esto no siempre es posible.

Hay un amplio acuerdo en que la asignación aleatoria del tratamiento es el mejor método para garantizar que los grupos sean inicialmente comparables respecto a todas las características pronósticas determinantes. Cuando es posible, conviene emplear enmascaramiento, es decir, que el paciente, la persona que aplica la intervención y la que evalúa el resultado desconozcan qué pacientes reciben el tratamiento nuevo y cuáles el tratamiento habitual (o el tratamiento control). Finalizado el seguimiento de los pacientes se procede a comparar los resultados de la intervención entre los grupos según su asignación aleatoria, lo que se denomina *el análisis por tratamiento asignado o por intención de tratar*.

Los estudios experimentales sin grupo control o sin un reparto aleatorio de la intervención entre los distintos grupos del estudio son muy susceptibles a incurrir en sesgos, por lo que sus resultados deben interpretarse con precaución.

Dentro de los ensayos controlados y aleatorizados, las unidades de asignación aleatoria pueden ser los individuos o los grupos, como sucede en los ensayos por conglomerados (ver capítulo de ensayos por conglomerados).

Por razones éticas estos estudios únicamente tienen aplicación para evaluar el efecto de exposiciones potencialmente terapéuticas o preventivas y, en menor medida, diagnósticas.

Ensayos clínicos

En los ensayos clínicos, también llamados ensayos terapéuticos, las unidades de asignación aleatoria suelen ser pacientes, aunque también puede haber ensayos clínicos por conglomerados. Su objetivo suele concretarse en evaluar nuevos tratamientos o estrategias de prevención secundaria. Sobre la población diana se aplican los criterios de inclusión y exclusión, previamente protocolizados, y se obtienen los sujetos elegibles.

Ensayos preventivos

■ Ensayos de campo

Los ensayos de campo evalúan medidas de prevención primaria y, por tanto, se aplican a sujetos sanos. Como la incidencia de la mayoría de las enfermedades en la población sana es muy baja, se requiere que estos estudios tengan un elevado número de participantes, para poder así obtener resultados concluyentes. Esto limita la aplicabilidad de los ensayos de campo a medidas de prevención en enfermedades frecuentes o muy graves y hace que sean estudios económicamente muy costosos. Un ejemplo sería la evaluación de las vacunas frente a la COVID-19 en la población sana para prevenir las complicaciones graves de la infección.

■ Ensayos comunitarios

Los ensayos comunitarios son ensayos por conglomerados que evalúan medidas de prevención primaria en grupos de gran tamaño, como barrios o ciudades (comunidades). En ellos, la aplicación de la intervención puede ser a nivel grupal o individual, mientras que la unidad de análisis puede ser el grupo, el individuo o ambos. Algunos ejemplos de posibles intervenciones comunitarias son la fluoración de las aguas de abastecimiento público, los programas de educación utilizando medios de comunicación de masas, o la intervención medioambiental en el lugar de trabajo. Cuando la asignación de la intervención no se hace de forma aleatoria entre los grupos o comunidades participantes, estamos ante estudios cuasiexperimentales.

3.4.2. Estudios cuasiexperimentales

Se utiliza el término *estudios cuasiexperimentales* para referirse a aquellos en los que hay una intervención no controlada (sin grupo control) o una asignación controlada de la exposición (intervención) sin seguir un procedimiento aleatorio. Las unidades de asignación del factor de estudio pueden ser los individuos o también las poblaciones. Cuando no se dispone de un grupo de control, la comparación se establece entre los mismos sujetos antes y después de aplicar la intervención.

Este tipo de estudios se utilizan frecuentemente para la evaluación de la eficacia de intervenciones poblacionales de tipo preventivo y de educación para la salud.

4. Estudios epidemiológicos y grado de evidencia científica

Rara vez se atribuye a un único estudio la evidencia científica definitiva respecto a un tema; sin embargo, son muchos los estudios epidemiológicos que aportan algún grado de evidencia a favor o en contra de presuntas asociaciones de tipo causal. Aun presumiendo la validez de los estudios, no todos aportan un mismo grado de evidencia. Según Maclure, (5) las características de los estudios epidemiológicos que permiten alcanzar una mayor validez son por orden de importancia: la aleatorización y el enmascaramiento, la desagregación del análisis a nivel del individuo, el diseño longitudinal y el planteamiento prospectivo. Por tanto, estas características en los estudios les conferirían un mayor grado de evidencia científica *a priori*.

En la tabla 3.6. se ordenan los principales diseños en función del grado de evidencia que pueden proporcionar. A los ensayos controlados y aleatorizados con enmascaramiento se les atribuye el mayor grado de evidencia científica dentro de la investigación epidemiológica. El resto de los ensayos y los estudios de cohortes, siempre que los métodos se hayan aplicado correctamente y se descarten sesgos importantes, se les reconoce el muy alto grado de evidencia. A los estudios de casos y controles y a los estudios transversales, cuando se han descartado sesgos importantes, se les atribuye un grado de evidencia elevado. Por último, a los estudios proporcionales, ecológicos y descriptivos en general, se les atribuye un valor orientativo, pero no concluyente.

Tabla 3.6. Clasificación de los diseños epidemiológicos según el grado de evidencia científica que se les atribuye en la investigación, ordenados de mayor a menor grado

1. Ensayo controlado y aleatorizado con enmascaramiento
2. Ensayo controlado y aleatorizado sin enmascaramiento
3. Ensayo controlado no aleatorizado, estudio de intervención
4. Estudio de cohortes concurrentes (prospectivas)
5. Estudio de cohortes históricas (retrospectivas)
6. Estudio de casos y controles con casos incidentes
7. Estudio de casos y controles con casos prevalentes
8. Estudio transversal
9. Estudio de mortalidad o morbilidad proporcional
10. Estudio ecológico de correlación

5. Referencias

1. Kramer MS, Boivin JF. Toward an “unconfounded” classification of epidemiologic research design. *J Chronic Dis*. 1987;40:683-8. doi: [10.1016/0021-9681\(87\)90104-4](https://doi.org/10.1016/0021-9681(87)90104-4).
2. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Principles and quantitative methods. Nueva York: Van Nostrand Reinhold; 1982.
3. Pearce N. Classification of epidemiological study designs. *Int J Epidemiol* 2012;41:393-7. doi: [10.1093/ije/dys049](https://doi.org/10.1093/ije/dys049).
4. Rothman KJ, Greenland S, Lash TL. Types of epidemiologic studies. En: Rothman KJ, Greenland S, Lash TL, editors. *Modern epidemiology*. Tercera edición. Filadelfia: Lippincott Williams & Wilkins; 2008. Páginas 87-99.
5. Maclure M. Taxonomic axes of epidemiologic study designs: a refutational perspective. *J Clin Epidemiol*. 1991;44:1045-53. doi: [10.1016/0895-4356\(91\)90006-u](https://doi.org/10.1016/0895-4356(91)90006-u).

Capítulo 4.

Sesgos

GREGORIO BARRIO, MARÍA JOSÉ BELZA y BELÉN SANZ BARBERO

1. Validez

1.1. Validez interna

1.2. Validez externa

2. Sesgos

2.1. Concepto e importancia

2.2. Tipos de sesgos

2.3. Sesgos de selección

2.4. Control de sesgos de selección

2.5. Sesgos de información o de clasificación errónea

2.6. Control de los sesgos de información

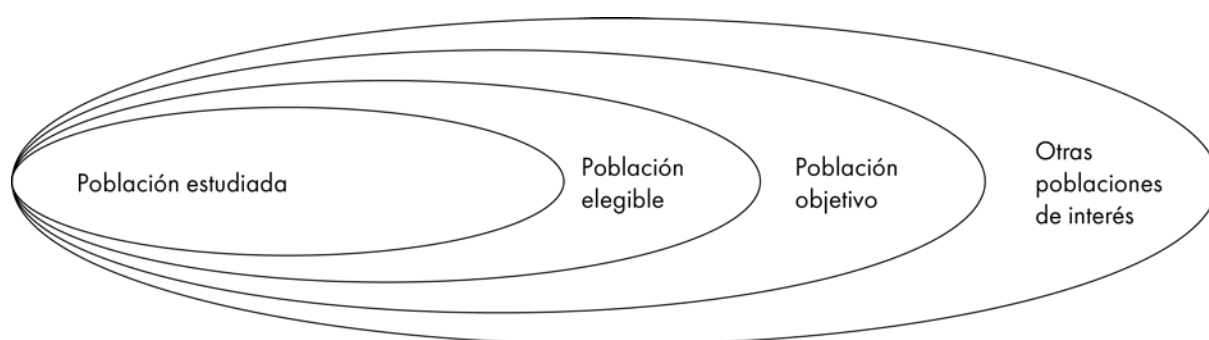
2.7. Otros sesgos

3. Referencias

1. Validez

La validez es el grado en que los resultados de un estudio epidemiológico pueden aplicarse a una población de interés especificada. En los textos de epidemiología pueden apreciarse inconsistencias a la hora de definir validez, que se trasladan también a la definición de sesgo, derivadas sobre todo de las distintas poblaciones a las que se pueden aplicar los resultados de un estudio (1). A continuación, se caracterizan estas poblaciones (Figura 1).

Figura 1. Tipos de población con implicaciones para estimar la validez del estudio y valorar los sesgos.



La “población estudiada” (participantes) son los sujetos en los que se miden u observan los parámetros de interés; la “población elegible” (marco muestral o población fuente) es la utilizada en la práctica para seleccionar a los participantes; la “población objetivo” es a la que se quieren aplicar los resultados del estudio; y las “otras poblaciones de interés” son poblaciones más amplias o diferentes de la objetivo (ubicadas en el mismo u otros lugares, momentos o entornos) a las que a veces se intentan generalizar los resultados. No siempre se pueden diferenciar los cuatro tipos de poblaciones. A menudo los resultados no se quieren generalizar más allá de una población objetivo concreta y otras veces coinciden la población elegible y la objetivo. En la Encuesta Nacional de Salud (estudio transversal) en la que se suele entrevistar a una muestra de 20 000-30 000 personas de 16 y más años residentes en viviendas familiares en España, seleccionadas mediante muestreo probabilístico trietápico, la población estudiada sería la muestra entrevistada, la población elegible serían las personas de 16 años que residen en viviendas familiares en España, y la población objetivo los residentes en España de ≥ 16 años, y en principio no se consideran otras poblaciones de interés porque no se quieren generalizar los resultados más allá de España y del año de la encuesta. En el estudio de seguimiento de Doll y Hill sobre la relación entre consumo de tabaco y cáncer

de pulmón (2) la población estudiada fue la muestra de médicos británicos analizada, la población elegible, los médicos registrados en las listas de las que se extrajo la muestra, la población objetivo el conjunto de médicos británicos, y las poblaciones de interés podrían ser todos los médicos o todos los adultos del mundo. No siempre es fácil definir la población objetivo y en muchos estudios no se define con claridad, lo que dificulta enormemente la valoración de los sesgos.

Los estudios epidemiológicos pretenden estimar uno o más parámetros con el menor error posible (máxima exactitud y precisión) en la población objetivo. En un estudio descriptivo suele ser la frecuencia (prevalencia o incidencia) de una enfermedad o una exposición, y en un estudio analítico (como ensayos controlados y aleatorizados, casos y controles, cohortes o series temporales) una medida de asociación entre exposición y enfermedad, como la *odds ratio* (OR) o la razón de riesgo (RR). Tal estimación puede estar afectada por errores aleatorios o sistemáticos.

El *error aleatorio* es la desviación del valor observado con respecto al valor real de la población objetivo debida al azar, por lo que desaparecería si la medición se repitiera un número infinito de veces en sujetos de esa población con igual metodología y en condiciones equivalentes, y se tomara el valor medio de las estimaciones. Las medidas repetidas estarían unas veces por encima y otras por debajo del valor real, y a distancia variable. Además, la desviación entre la media de las estimaciones y el valor real (error aleatorio) disminuirá al aumentar el número de repeticiones. El error aleatorio se debe sobre todo al error muestral derivado de estudiar solo una parte de la población objetivo. Otras fuentes de error aleatorio pueden ser la imprecisión del instrumento de medida o variaciones imprevisibles del parámetro estimado. Por ejemplo, si en un estudio de prevalencia de hipertensión arterial en cierta población se extrae una muestra aleatoria con una prevalencia del 20 % y otra muestra aleatoria de igual tamaño con una prevalencia del 22 %, y se ha usado la misma metodología y no hay errores por factores ajenos al azar, la diferencia se debería al error aleatorio y disminuiría al aumentar el tamaño muestral. Un error aleatorio alto implica baja precisión o reproducibilidad de los resultados. El error aleatorio dificulta la detección y caracterización de las asociaciones entre variables. En su valoración influye sobre todo el tamaño muestral, aunque también deben considerarse elementos de juicio informado.

El *error sistemático o sesgo* es la desviación del valor observado con respecto al valor real de la población objetivo debida a factores distintos al azar. Es constante en cada medición y va siempre en el mismo sentido (por encima o por debajo del valor real). Al contrario que el aleatorio, no desaparece al repetir el estudio muchas veces con igual metodología y su magnitud persiste al aumentar el tamaño de la muestra. En su valoración intervienen fundamentalmente elementos de juicio informado (Tabla 1).

Tabla 1. Principales diferencias entre error aleatorio y sistemático

	ERROR ALEATORIO	ERROR SISTEMÁTICO (SESGO)
Causa	Inherente al muestreo. Variaciones imprevisibles del observador, instrumento de medida o parámetro medido.	Factores ajenos al muestreo, relacionados con deficiencias en el diseño, ejecución, análisis, interpretación o publicación del estudio.
Sentido	La dirección del error no es identificable, por definición.	En un único sentido.
Valoración	Métodos estadísticos. Juicio razonado.	Juicio razonado.
Control	No puede evitarse. Puede minimizarse con métodos estadísticos, aumentando el tamaño muestral.	Excepto el debido a confusión, no puede controlarse con métodos estadísticos. Puede reducirse con buenas prácticas en el diseño, ejecución, interpretación y publicación del estudio.

La validez tiene dos componentes, la validez interna y la validez externa, aunque la línea de separación entre ambas depende de la población a que se quieran aplicar los resultados. (1)

1.1. Validez interna

Es el grado en que los resultados observados son ciertos para la población objetivo y no se explican por errores sistemáticos (sesgos) o confusión debidos a deficiencias en el diseño, ejecución, análisis, interpretación o publicación del estudio. En los estudios que investigan relaciones causales es el grado en que los cambios observados en el desenlace al variar el nivel de exposición pueden explicarse por el efecto de dicha exposición y no por otros factores. Así, en el estudio de seguimiento de Doll y Hill citado, (2) los resultados tendrán validez interna si la relación entre tabaco y cáncer de pulmón encontrada en los médicos estudiados es cierta en esos médicos y en todos los médicos británicos (población objetivo). Todos los estudios epidemiológicos tienen problemas de validez interna en mayor o menor grado ya que es imposible controlar todas las posibles fuentes de sesgos y de confusión.

1.2. Validez externa

Es el grado en que los resultados del estudio son aplicables a poblaciones de interés más amplias o diferentes a la población objetivo, tanto en los mismos como en otros lugares, momentos o circunstancias. Es, por tanto, la capacidad de generalizar los resultados más allá de la población objetivo. En el estudio de Doll y Hill citado, (2) los resultados tendrán validez externa si la relación encontrada en médicos británicos es cierta en toda la población inglesa o de otros sitios o si el consumo de tabaco aumenta el riesgo de cáncer de pulmón en cualquier grupo de fumadores.

La validez interna es un prerequisite de la externa. Esta última puede estar comprometida si las condiciones de investigación no son las mismas que las de la población de interés o hay variables relevantes de esta población relacionadas con las variables del estudio que no se han considerado. La validez externa es una cuestión de juicio informado y para valorarla hay que recurrir a criterios e información externas al estudio. Para muestreos no representativos, en principio, la generalización de resultados se limita a poblaciones con características similares a la muestra. Los resultados de ciertos estudios pueden tener alta validez externa para toda la población mundial, mientras que otros podrían tenerla solo para ciertas poblaciones de interés, por ejemplo, no se puede generalizar más allá de la población objetivo los estudios dirigidos a estimar la prevalencia de condiciones de salud, pero las relaciones causales son con frecuencia generalizables, sobre todo cuando hay razones para pensar que los mecanismos causales, conocidos o postulados, son similares en diferentes poblaciones (el tabaco produce efectos parecidos en poblaciones de Europa, Asia o África, ya que todas pertenecen a la misma especie, con mínimas variaciones, genéticas, biológicas o fisiológicas; sin embargo, la relación de la posición socioeconómica con la obesidad varía con las circunstancias del entorno, como la disponibilidad y precio de diferentes tipos de alimentos o el producto interior bruto del país de residencia).

2. Sesgos

2.1. Concepto e importancia

Un sesgo es un error sistemático en los resultados de un estudio con respecto a los reales en la población objetivo. Limita la validez interna del estudio (es decir,

la validez de las inferencias a la población objetivo), y al hacerlo limita también su validez externa (es decir, la validez de las inferencias fuera o más allá de la población objetivo). Afecta a los resultados siempre en un sentido identificable, sobreestimando o subestimando los valores reales.

La repercusión relativa del sesgo es mayor si el efecto real de la exposición es débil o de pequeña magnitud porque puede llegar a invertir su sentido y hacer que un factor de riesgo parezca protector o al revés. En cambio, si el efecto real es fuerte, es más difícil que el sesgo pueda anularlo o invertir su sentido. Como cada estudio suele estar sujeto a sesgos diferentes, si se detecta un efecto fuerte y homogéneo en la mayoría de los estudios es muy probable que se dé realmente en la población objetivo. La relevancia de los sesgos ha aumentado al desarrollarse las técnicas de síntesis de los resultados científicos mediante revisiones sistemáticas y metanálisis. Así, la herramienta GRADE para valorar la calidad de los resultados científicos y la fuerza de las recomendaciones incluye como primer criterio para calificar esa calidad el riesgo de sesgo de los estudios individuales de base. (3,4,5)

Los sesgos se deben a deficiencias al diseñar, ejecutar, analizar, interpretar o publicar el estudio, con frecuencia por malas prácticas. La susceptibilidad a sesgos depende mucho del tipo de estudio epidemiológico. Por ejemplo, en los ensayos controlados y aleatorizados la asignación aleatoria de la intervención controla, si la muestra es suficientemente grande, los factores de confusión medidos y los no medidos, por lo que la susceptibilidad a sesgos de confusión es mínima. Y dentro de los estudios observacionales, los estudios transversales y de casos y controles son más proclives a sesgos de selección que los de cohortes, porque en los primeros la probabilidad de seleccionar enfermos y no enfermos puede estar influida por la exposición, mientras que en los estudios de cohortes esta probabilidad de selección diferencial es menor porque la exposición se mide antes de aparecer la enfermedad. Además, con un mismo diseño epidemiológico, las medidas subjetivas suelen ser más proclives a sesgos de información, incluyendo el de memoria o recuerdo, que las objetivas. Por otra parte, en estudios de cohortes ciertas medidas subjetivas (ej.: calidad de vida), son mucho más susceptibles a sesgos por pérdida de seguimiento que la mortalidad por todas las causas, algo que a veces se ignora en las revisiones sistemáticas.

Todos los estudios epidemiológicos, sobre todo los observacionales, son proclives a sesgos, aunque pocas veces se tendrá certeza de su magnitud porque se desconoce el valor real. En los estudios epidemiológicos analíticos es prácticamente imposible lograr grupos de comparación idénticos que solo se diferencien en la exposición de interés como sucede en los experimentos con animales genéticamente similares. Por eso, en cada estudio hay que intentar argumentar razonadamente sobre la presencia e influencia de posibles sesgos en los resultados, algo que suele

hacerse en la sección de discusión de los artículos científicos. Por ejemplo, se sabe que al autoinformar sobre nivel de ejercicio físico se tiende a sobrestimarlo, por lo que pueden hacerse conjeturas sobre la magnitud del sesgo. Sin embargo, la existencia de varios sesgos en el mismo estudio dificulta la estimación de la influencia conjunta. Siempre ha de valorarse la influencia de los sesgos en los resultados de un estudio epidemiológico, y se recomienda un análisis cuantitativo formal del sesgo cuando los resultados del estudio vayan a usarse para inferencias causales, recomendar intervenciones o tomar decisiones relevantes, ya que permite hacerse una idea de la incertidumbre derivada de los errores sistemáticos y contrarrestar el exceso de confianza en los resultados de la investigación y las inferencias basadas en ellos. (6)

A diferencia del error aleatorio, los sesgos (excepto el de confusión) en general no pueden controlarse con métodos estadísticos, aunque pueden minimizarse con buenas prácticas al diseñar, ejecutar, interpretar y publicar el estudio, especialmente mediante una adecuada selección de participantes y medida de las variables.

2.2. Tipos de sesgos

Hay muchos tipos de sesgos y diferentes clasificaciones. A menudo se utilizan distintos términos o etiquetas para referirse al mismo fenómeno, lo que dificulta la comunicación entre epidemiólogos y otros profesionales. (1)

Los textos de epidemiología suelen distinguir tres tipos de sesgos: 1) de selección, por falta de representatividad de los participantes, 2) de información, por información inadecuada de los participantes, y 3) de confusión o distorsión del efecto por terceras variables no controladas. Sin embargo, no siempre es fácil clasificarlos en estos tres grupos. Algunos tienen a la vez componentes de selección e información. Así, un mismo sesgo puede concebirse como de selección en un tipo de diseño epidemiológico (ej.: casos y controles) y de información en otro (ej.: cohortes). Además, hay factores que pueden generar sesgos de selección que pueden considerarse factores de confusión. Algunos autores consideran que un sesgo es de confusión si puede medirse y controlarse con técnicas estadísticas al analizar los datos. Sea cual sea el tipo de sesgo, en los estudios analíticos el resultado es la falta de comparabilidad entre los grupos definidos por la exposición de interés.

Finalmente hay sesgos que no son fáciles de integrar en la clasificación anterior, como el de causalidad inversa o la falacia ecológica, que se ponen de manifiesto al interpretar los resultados, o el sesgo de publicación y relacionados.

No obstante, más importante que nombrar o clasificar un sesgo es entender su origen e influencia en los resultados en cada estudio concreto. (1,7)

2.3. Sesgos de selección

Son errores sistemáticos derivados de que los sujetos estudiados o participantes no son representativos de la población objetivo, bien por muestreo no aleatorio, en estudios de prevalencia, o bien porque los sujetos de dicha población tienen distinta probabilidad de participar según estén en una u otra categoría de exposición o desenlace, en estudios analíticos. Se deben a malas prácticas al seleccionar a los participantes o a factores que condicionan la participación en el estudio. Esto hace que los participantes no sean comparables con los no participantes pertenecientes a la población objetivo, por lo que el valor del parámetro estimado, por ejemplo, la razón de riesgo, diferirá del que se obtendría si se estudiara toda la población objetivo. Como las estimaciones del efecto están condicionadas por la participación, las asociaciones observadas son una mezcla de los factores que determinan la participación y los que determinan la aparición de la enfermedad.

El sesgo de selección limita la validez interna. En estudios analíticos es más probable que aparezca cuando los grupos de expuestos y no expuestos (estudios de cohortes) o de enfermos y no enfermos (estudios de casos y controles) no se extraen de la misma población. Por ejemplo, en un estudio sobre la actividad física materna y problemas en el parto, las embarazadas expuestas (activas) se reclutaron entre las mujeres que acudían a clases de ejercicios prenatales privadas y las no expuestas (sedentarias) en clínicas públicas de atención prenatal. Como las expuestas pagan y van voluntariamente a clase suelen ser diferentes a las no expuestas y los resultados pueden estar afectados por un sesgo de selección derivado de determinantes socioeconómicos.

Estos sesgos pueden introducirse al seleccionar los participantes (por mala definición de la población elegible, falta de delimitación del marco muestral, aplicación de pruebas diagnósticas de selección diferentes), al implementar el estudio (sobre todo por pérdidas de seguimiento, negativa a participar o falta de datos sobre ciertas variables) o al publicar o seleccionar los estudios para extraer conclusiones. (7,8)

2.3.1. Sesgos al seleccionar los participantes

Se producen en el momento de seleccionar los participantes antes de ejecutar o recoger información sobre las variables del estudio. Son muy frecuentes y a menudo indetectables cuando no se dispone de un marco muestral bien delimitado (listado de elegibles) y se hace un muestreo de conveniencia para seleccionar a los participantes. Los principales sesgos en esta fase son los siguientes:

- **Sesgo por muestreo no representativo:** Suele ser el sesgo más importante en los estudios transversales de prevalencia. Por su falta de representatividad,

el estudio genera estimaciones de los parámetros distintas a los valores en la población objetivo. Se produce, por ejemplo, en encuestas domiciliarias cuando se excluye del marco muestral a la población sin hogar o institucionalizada (residencias, prisiones, hospitales, etc.), o en encuestas en línea cuando se excluye a las personas sin conexión a internet. Es muy similar al sesgo por acceso a los servicios de salud, que se da en estudios de prevalencia o incidencia cuando los pacientes reclutados en instituciones sanitarias no representan a los de la población general. Este sesgo afecta solo a la validez externa, mientras que el resto de sesgos de selección descritos a continuación afectan tanto a la validez interna como a la externa. (9)

- **Sesgo de autoselección:** La aceptación o probabilidad de participar en el estudio se relaciona con la exposición o con el desenlace. Hay variantes. Cuando los participantes se autoseleccionan directamente para el estudio, se puede producir el sesgo del voluntario, porque los voluntarios suelen tener conocimientos, actitudes, conductas y estados de salud diferentes a los no voluntarios. Un ejemplo es el estudio de leucemia en militares expuestos a pruebas atómicas en Nevada. De todos los expuestos reclutados (76%), un 62% lo fueron por investigadores y un 14% se presentaron voluntarios tras recibir publicidad del estudio, encontrándose 4 leucemias en cada grupo, lo que supone una incidencia mucho más alta en voluntarios. (10)

La autoselección puede producirse antes de seleccionar a los participantes, como sucede en el sesgo del trabajador sano, en estudios observacionales sobre el efecto de exposiciones ocupacionales raras en la mortalidad que usan población general como grupo control no expuesto. El problema es que los empleados (sobre todo en trabajos que requieren un buen estado físico, como bomberos o militares) suelen tener menor mortalidad que la población general (que incluye también desempleados y personas con discapacidad), por lo que se subestima el efecto de la exposición. Supóngase que los empleados expuestos tienen una tasa de mortalidad ajustada por edad de 150/100 000 personas-año (p-a), los empleados no expuestos 75/100 000 p-a y la población general (en su gran mayoría sin la exposición rara) 100/100 000 p-a. Si se comparan los empleados expuestos con los no expuestos, la razón de tasas será 2,0, pero si se comparan con población general, será 1,5, subestimándose la magnitud real de la asociación. Este sesgo también puede considerarse como confusión, ya que se debe a un tercer factor o cúmulo de factores que se asocian tanto con el estatus de empleo como con la mortalidad. Un sesgo similar es el sesgo del migrante sano, que explicaría en parte por qué en países ricos los inmigrantes labora-

les de países pobres tienen tasas de mortalidad ajustadas por edad más bajas que los autóctonos, ya que migran para trabajar los más sanos. (7)

- Sesgo de incidencia-prevalencia o de *Neyman*: Es muy común. Se da cuando se estiman asociaciones potencialmente causales entre una exposición y la aparición de una enfermedad usando casos prevalentes (donde suelen estar sobrepresentados los de mejor pronóstico y menos expuestos) en vez de incidentes (donde en principio están representados proporcionalmente todos los niveles de gravedad). En estos estudios el efecto de la exposición se estima con la razón de prevalencias puntuales entre expuestos y no expuestos. Como la prevalencia, en determinadas condiciones, depende de la tasa de incidencia y de la duración media de la enfermedad (ver capítulo 2), el sesgo puede aparecer si la duración media de la enfermedad difiere en expuestos y no expuestos. Si la exposición se asocia con la duración de la enfermedad, se denomina *sesgo de supervivencia selectiva*, caracterizado generalmente por una subestimación de la razón de riesgo, ya que los casos expuestos suelen tener menor supervivencia (menor duración de la enfermedad). Sin embargo, la razón de prevalencias también puede subestimar la razón de riesgo si las prevalencias son altas. La demostración algebraica de estos sesgos se muestra en el capítulo 2.4 sobre medida del efecto en estudios de prevalencia (transversales).

Los estudios transversales y de casos y controles son los más susceptibles a este sesgo, aunque también puede darse en los experimentales y de cohortes. A veces, en los estudios de casos y controles que intentan reclutar casos incidentes, este sesgo aparece sin que se percate el investigador porque hay enfermedades que pueden evolucionar subclínicamente mucho tiempo antes del diagnóstico y los supuestos casos incidentes seleccionados son en realidad una mezcla de casos incidentes y prevalentes. Para ilustrar el sesgo imagínese un estudio de casos y controles en el que se reclutan todos los casos incidentes en un periodo (400), de los que la mitad son muy graves y mueren muy rápido (25 % expuestos) y el resto menos graves y sobreviven (10 % expuestos) y 400 controles (5 % expuestos). La Tabla 2 muestra que el efecto de la exposición estimado es menor para casos prevalentes ($OR=2,1$) que para casos incidentes ($OR=4,0$).

Tabla 2. Ejemplo de estimación del efecto de una exposición para casos incidentes y prevalentes

	Expuestos	No expuestos	Total	Efecto (OR)
Casos incidentes	70	330	400	4,03
Controles	20	380	400	
Total	90	710	800	
	Expuestos	No expuestos	Total	Efecto (OR)
Casos prevalentes	40	360	400	2,11
Controles	20	380	400	
Total	60	740	800	

OR: Odds Ratio

Por lo tanto, el estudio de casos prevalentes subestima la asociación en comparación con el de casos incidentes. Sin embargo, si la exposición fuera un factor de buen pronóstico y mayor supervivencia, un estudio de casos prevalentes aumentaría la probabilidad de seleccionar expuestos y se sobreestimaría la OR. (8)

- Sesgo de los pacientes hospitalizados o de *Berkson*: Se produce en estudios transversales o de casos y controles que seleccionan a los participantes entre personas ingresadas en hospitales, cuando la probabilidad de ingreso hospitalario está afectada tanto por la exposición como por la enfermedad; es decir, difiere para casos y controles y según la exposición. Los pacientes hospitalizados suelen diferir en muchos aspectos de la población general de la que surgen, sobre todo en cuanto la probabilidad de exposición a los factores de riesgo estudiados, por ejemplo, consumo de tabaco o alcohol. Aun así, muchas veces se recurre a ellos como casos y también como controles por razones prácticas, ya que son más fáciles de conseguir que los controles comunitarios.

Las razones de su aparición pueden ser diferentes en cada hospital o centro concreto, dependiendo de su política de ingresos, facilidad de acceso, servicios que ofrece o popularidad. Puede dar lugar a subestimación (lo más frecuente) o sobrestimación del efecto de la exposición cuando se analiza un factor de riesgo de la enfermedad. Por ejemplo, en un estudio sobre el efecto de fumar sobre el cáncer de pulmón los casos fueron los pacientes con este cáncer ingresados en un hospital y los controles los ingresados sin

dicho cáncer. Sin embargo, los fumadores tienen mayor probabilidad de ingreso por patologías diferentes al cáncer de pulmón que los no fumadores. Por ello, la prevalencia de fumar entre los controles será mayor que en la población general de la que surgen los casos (población objetivo) y la OR estará sesgada hacia el valor nulo. En otro estudio sobre el efecto del café sobre el cáncer de páncreas los casos fueron pacientes con este cáncer hospitalizados y los controles pacientes con enfermedades digestivas del mismo hospital. Estos últimos tenían menor probabilidad de tomar café que el resto de la población debido a su enfermedad digestiva, por lo que se sobrestimó la asociación entre tomar café y cáncer de páncreas. (7,11,12)

- **Sesgo de sospecha diagnóstica o de detección:** Se da en los estudios de casos y controles cuando la probabilidad de selección de los casos está influida por la presencia o el conocimiento de la exposición. En general produce una sobreestimación del efecto, especialmente si la exposición es un factor de riesgo conocido o un tratamiento que conduce a una vigilancia médica más intensa o a pruebas diagnósticas más frecuentes. En estas circunstancias, los expuestos tendrán mayor probabilidad de que se les diagnostique la enfermedad y de ser seleccionados como casos, sobre todo si el estudio se basa en casos prevalentes y la enfermedad tiene un importante componente subclínico. Así, podría observarse una asociación espuria entre uso de contraceptivos orales y fallo renal, si las mujeres que toman contraceptivos se hicieran con más frecuencia determinaciones de creatinina sérica. Sin embargo, podría producir una subestimación del efecto cuando la probabilidad de revisiones médicas sea mayor en los casos no expuestos que en los expuestos, como ocurrió en un estudio de cohortes sobre tabaquismo y basalioma, en el que se observó un menor riesgo de este cáncer en fumadores que en no fumadores, probablemente porque los primeros se hacían menos revisiones médicas cutáneas. Este sesgo también se da en estudios de cohortes, donde puede concebirse como un sesgo de información, (7,8,12) como se ilustrará en la sección correspondiente a los sesgos de información.
- **Sesgo de inclusión-exclusión:** Suele ocurrir en los estudios de casos y controles cuando se incluyen o se excluyen sistemáticamente otras enfermedades o condiciones relacionadas con la exposición estudiada solo en uno de los dos grupos, o en los casos o en los controles. Esto equivale a decir que los criterios de selección de participantes están afectados por la exposición estudiada. Si se incluyen condiciones relacionadas con la exposición en los controles y no en los casos, aumenta su probabilidad de exposición con respecto a la población objetivo y se subestima el efecto, y lo contrario ocurre

si se excluyen condiciones con características similares. Este sesgo se produce porque muchas exposiciones son factores de riesgo de varias enfermedades. Un ejemplo clásico es la detección de una asociación espuria entre toma del antihipertensivo reserpina y cáncer de mama. En un estudio se comparó su consumo en mujeres con cáncer de mama con el de mujeres ingresadas para cirugía de procesos benignos (controles). Del grupo control se excluyeron trastornos que en aquel momento se trataban con reserpina como tiroidectomía por tirotoxicosis o cirugía por enfermedad renal o cardíaca, un criterio que no se aplicó a los casos, por lo que se subestimó el consumo de reserpina en los controles con respecto a la población objetivo de la que provenían los casos y se sobreestimó la medida de asociación. (7,13)

- **Sesgo de numerador-denominador:** Es un sesgo bastante frecuente, pero poco citado en los manuales de epidemiología. Puede considerarse también un sesgo de información. Suele aparecer en los estudios de morbilidad no vinculados cuando los datos del numerador (muertes o casos de enfermedad) y denominador (población residente) proceden de fuentes diferentes. En estas circunstancias, si se quiere estudiar la asociación entre un factor de riesgo (ej.: nivel de estudios) y un desenlace (ej.: mortalidad relacionada con alcohol), los resultados pueden estar sesgados si el nivel educativo se ha clasificado de forma diferente en los registros de morbilidad (numerador) y de población (denominador). (14)

2.3.2. Sesgos al ejecutar el estudio

Se producen una vez seleccionados los participantes e iniciado el estudio. Aunque los participantes se hayan seleccionado adecuadamente, durante la ejecución del estudio pueden surgir problemas que producen un efecto sobre la muestra final analizada similar al de una selección inicial inadecuada. Los principales sesgos de selección en esta fase son el de pérdida de seguimiento y el de no respuesta, que son los sesgos más relevantes en los estudios de seguimiento (cohortes y ensayos controlados y aleatorizados). Estos sesgos se deben a la ausencia de información sobre el desenlace u otras variables de interés en parte de los sujetos seleccionados para el estudio (censura de los datos).

- **Sesgo por pérdida de seguimiento:** Se produce si algunos participantes salen prematuramente del estudio (pérdidas, ausencias o abandonos) de forma diferencial, es decir, si su salida está relacionada con la exposición o con la enfermedad. En estas circunstancias la incidencia de enfermedad o la probabilidad de exposición en la cohorte diezmada será diferente de la que hubiese existido en la cohorte completa y se verá comprometida la estimación de la asociación real existente en la población objetivo. Si las pérdidas

fueran al azar y la cohorte suficientemente grande, afectarían por igual a los cuatro subgrupos resultantes de combinar las categorías de exposición y enfermedad, y no se produciría sesgo al estimar la razón de riesgo. Tampoco se produciría si las pérdidas afectaran de forma desigual a expuestos y no expuestos, pero en unos y otros se mantuviera la misma proporción en enfermos y en no enfermos, o, de forma general, si afectaran de forma desigual a enfermos y no enfermos, pero en unos y otros se mantuviera la misma proporción en expuestos y no expuestos. No obstante, Greenland demostró que es suficiente con que las pérdidas se asocien con el desenlace, aunque no lo hagan con la exposición, para que se produzca sesgo de selección cuando la exposición se asocia con la enfermedad en la población objetivo (ver capítulo 7). (15) Para que haya sesgo las pérdidas han de ser diferenciales; es decir, han de relacionarse al menos con la enfermedad. En la Tabla 3 se ejemplifican numéricamente las distintas situaciones en una cohorte de 1000 personas expuestas y 1000 no expuestas seguidas durante cinco años para observar la aparición de una enfermedad.

Tabla 3. Efectos de la pérdida diferencial de seguimiento sobre la estimación de la asociación en un estudio de cohortes. Estudio ficticio

	I. Sin pérdidas de seguimiento			
	Expuestos	No expuestos	Total	Efecto (RR)
Enfermos	30	10	40	3,0
No enfermos	970	990	1960	
Total	1000	1000	2000	
	II. Con un 20 % de pérdidas al azar (se espera que afecten en igual proporción a todos los subgrupos)			
	Expuestos	No expuestos	Total	Efecto (RR)
Enfermos	24	8	32	3,0
No enfermos	776	792	1568	
Total	800	800	1600	

Tabla 3. Efectos de la pérdida diferencial de seguimiento sobre la estimación de la asociación en un estudio de cohortes. Estudio ficticio (continuación)

	III. Con pérdida no diferencial a los 5 años de 20 % en enfermos y 10 % en no enfermos			
	Expuestos	No expuestos	Total	Efecto (RR)
Enfermos	24	8	32	3,0
No enfermos	873	891	1764	
Total	897	899	1796	
	IV. Con pérdida no diferencial a los 5 años de 20 % en expuestos y 10 % en no expuestos			
	Expuestos	No expuestos	Total	Efecto (RR)
Enfermos	24	9	33	3,0
No enfermos	776	891	1667	
Total	800	900	1700	
	V. Con pérdida diferencial a los 5 años de 20 % en expuestos y 10 % en no expuestos			
	Expuestos	No expuestos	Total	Efecto (RR)
Enfermos	29	8	37	4,1
No enfermos	771	892	1663	
Total	800	900	1700	

RR: Razón de Riesgo.

El sesgo por pérdida de seguimiento puede ir en cualquier dirección. Cuanto mayor es la duración del período de observación, más difícil es seguir a todos los participantes. Las pérdidas pueden producirse por muerte, cambio de domicilio o pérdida de interés en el estudio, entre otros motivos. Por ejemplo, en estudios de adultos mayores, los pacientes más graves tienden a abandonar porque la enfermedad les dificulta asistir a las visitas clínicas del estudio, lo que puede dar lugar a sesgo de selección. También puede aparecer en estudios de mortalidad laboral, cuando, por ejemplo, la exposición hace que algunos participantes abandonen una ocupación prematuramente, lo que puede impedir o dificultar localizarlos, aumentando la probabilidad de excluirlos del estudio. Como estas personas suelen tener mayor riesgo de en-

fermedad y muerte motivada por la exposición, hay menor probabilidad de selección de los casos expuestos y una subestimación del efecto.

Las pérdidas de seguimiento son más frecuentes en participantes captados en la comunidad que en pacientes de centros asistenciales, porque estos últimos se sienten más vinculados con los sanitarios que les han propuesto participar. Un sesgo similar al de pérdida de seguimiento es el sesgo por *riesgos competitivos*, que se puede dar si hay otro u otros desenlaces que compiten con el de interés en el mismo sujeto y son mutuamente excluyentes. Así sucede cuando el desenlace es la muerte por una causa dada y el riesgo de muerte por esta causa resulta afectado por la mortalidad previa por otras causas. Por ejemplo, en un estudio de seguimiento de inyectadores de drogas el riesgo de muerte por VIH en expuestos y no expuestos a una intervención podría verse afectado de forma diferencial por la mortalidad prematura por sobredosis. Otro sesgo que tiene similitudes con el de pérdidas diferenciales, aunque se da también en estudios transversales es el sesgo del *salmón*, que se produce cuando los inmigrantes muy enfermos retornan para morir al país de origen y no se computan en las estadísticas de mortalidad del país receptor. (3,8)

- Sesgo de no respuesta: Se debe a la no participación de algunos seleccionados en el estudio o a la ausencia de valores en ciertas variables de interés (valores ausentes), si la no respuesta está relacionada con la exposición o la enfermedad. Puede deberse a la imposibilidad de localizar al participante, o a que rechaza participar o rechaza ciertas preguntas, exámenes o análisis. En un estudio de casos y controles para analizar la relación entre fumar tabaco y enfermedad isquémica, se intentó minimizar el rechazo a participar entre los controles comunitarios. A los controles seleccionados se les invitó a acudir al centro de estudio para extraerles sangre, tomar medidas antropométricas y contestar un cuestionario sobre factores de riesgo. A los que no acudieron, se les envió un breve cuestionario para completar en casa. Por último, se acudió al domicilio de los que no contestaron el cuestionario breve. Se observó que, a mayor resistencia a participar, mayor probabilidad de ser fumador, especialmente entre mujeres. Suponiendo que el consumo de tabaco de los que rechazaron completamente participar fuera similar o mayor que el de los reacios a participar, la prevalencia de fumar estaría subestimada con respecto a la población objetivo y la asociación sobreestimada. En general, las personas con peor salud y conductas menos saludables participan menos en encuestas de salud. (12,13)
- Sesgo de colisión (*collider bias*): Se usa este término para referirse a la detección de una asociación espuria entre una exposición y un desenlace cuan-

do ambas presentan una relación causal con un efecto común, el factor de colisión, y esta variable se controla en el diseño del estudio (restricción) o mediante análisis estadístico. Puede considerarse un sesgo de selección porque supone que los participantes del estudio se identifican de tal manera que ya no son representativos de la población objetivo. (16) Los sesgos de selección (inclusión/exclusión, Berkson o pérdida de seguimiento diferencial) pueden concebirse y representarse gráficamente como ejemplos de sesgo de colisión.

2.4. Control de sesgos de selección

Una estrategia general para evitar o minimizar los sesgos de selección es procurar seleccionar a los participantes de una población objetivo bien definida y que en los estudios retrospectivos los casos y los controles estén sometidos al mismo proceso de selección. Son ejemplos de buena práctica en este sentido los estudios de casos y controles que, para investigar los factores de riesgo del cáncer de mama, reclutan tanto a los casos como a los controles entre mujeres que participan en los programas de cribado de este cáncer. Las mujeres que participan en estos programas suelen tener mayor probabilidad de factores de riesgo conocidos (ej: antecedentes familiares) y si los controles se reclutaran entre mujeres de la población general, podría producirse una sobreestimación de la asociación con algunos factores. Al seleccionar tanto los casos como los controles en los programas de cribado se espera que ambos grupos tengan el mismo nivel de sobreexposición a los factores de riesgo y que las medidas de efecto sean insesgadas. En la práctica, sin embargo, es muy difícil asegurar que la sobrerrepresentación afecta por igual a casos y controles y a veces el deseo de lograrlo tiene efectos contraproducentes como sucedió en el ejemplo del cáncer de páncreas y consumo de café. No obstante, es necesario que los controles sean una muestra de una población objetivo especificada de la cual proceden también los casos. (8)

En los estudios transversales y de casos y controles, los sujetos a menudo se seleccionan después de producirse la exposición y el desenlace (enfermedad), por lo que hay un alto potencial de sesgo debido a los efectos combinados de la exposición y la enfermedad en la selección. Como los sesgos de selección pueden considerarse también como confusión provocada por factores que influyen en la selección antes de la exposición o del diagnóstico, una estrategia para minimizarlos o controlarlos puede ser identificar y medir de antemano tantos factores de este tipo como sea práctico, lo que requiere dinero y un buen conocimiento de la materia o del marco teórico. Sin embargo, en la práctica muchos de estos sesgos de selección no pueden controlarse con los métodos de ajuste de covariables clásicos, por lo que hay que extremar las medidas realizando una selección cuidadosa de

los participantes, procurando que sean lo más representativos posible de los casos aparecidos en la población objetivo tras la exposición. (7,12)

A continuación, se exponen estrategias para reducir el efecto de algunos sesgos de selección:

- **Sesgo por muestreo no representativo:** Utilizar siempre que sea posible muestreos probabilísticos de la población objetivo y controlar los procesos de selección, incluso cuando lo realizan empresas externas al propio equipo investigador, de forma que la muestra sea representativa. Cuando este sesgo se debe a que la población elegible (de la que se extrae la muestra) no coincide con la población objetivo (a la que se quieren aplicar los resultados), podría redefinirse la población objetivo y aplicar los resultados solo a la población elegible. (9) Algunas técnicas estadísticas, como la ponderación por el inverso de la probabilidad de selección pueden mitigar este sesgo, pero solo si todos los predictores relevantes del desenlace se miden e incluyen en la corrección. (17)
- **Sesgo de autoselección:** El sesgo del voluntario puede disminuir garantizando anonimato y confidencialidad, y evitando que los controles conozcan la hipótesis de estudio para que la participación no dependa de la misma. Otra estrategia es obtener la muestra de una cohorte preestablecida con altas tasas de participación. Además, si se recopila cierta información demográfica sobre los que consienten y no consienten participar, se pueden identificar predictores del no consentimiento y realizar una cierta corrección en el análisis mediante ponderación por probabilidad inversa. El sesgo del trabajador sano puede evitarse comparando los trabajadores expuestos con trabajadores no expuestos y no con población general. Se pueden seleccionar como no expuestos a trabajadores de la misma empresa o de otra que use procedimientos de reclutamiento laboral similares para que los dos grupos sean comparables con respecto a variables relevantes. Si el grupo de comparación es interno, puede comprobarse si existe un gradiente dosis-respuesta en el desenlace entre grupos de trabajadores con alto medio o bajo nivel de exposición. (7,17)
- **Sesgo de incidencia-prevalencia:** Se puede evitar seleccionando solo casos incidentes. Un estudio de cohortes puede prevenir bien este sesgo si se excluyen los casos prevalentes al inicio del seguimiento y hay una búsqueda sistemática y periódica de los nuevos casos, incluyendo los desenlaces subclínicos. (17)
- **Sesgo de Berkson:** una estrategia para minimizarlo o evitarlo es, como se ha señalado, procurar que los casos y los controles estén sometidos al mismo proceso de selección o, más concretamente, utilizar como controles pacientes que tengan una probabilidad de ingreso hospitalario similar a los

casos. Aunque es difícil conocer esa probabilidad de forma precisa, sí suele conocerse el rango de probabilidad de diferentes enfermedades. (18)

- Sesgo de sospecha diagnóstica: En los estudios de cohortes prospectivos se puede prevenir con dos estrategias: a) aplicando los mismos procedimientos diagnósticos para evaluar el desenlace (enfermedad) a expuestos y no expuestos, algo más factible cuando el estudio cuenta con sus propios procedimientos diagnósticos que cuando depende del sistema rutinario de atención médica, y b) enmascarando el estado de exposición al evaluar el desenlace (el evaluador desconoce la exposición del participante). Sin embargo, estas estrategias generalmente no se pueden aplicar en los estudios de casos y controles porque el diagnóstico de la enfermedad suele ser previo al inicio del estudio. En estas circunstancias cabe hacer un juicio sobre su presencia y magnitud basándose en información sobre la frecuencia y calidad de los procedimientos diagnósticos aplicados a los participantes y los posibles factores asociados como posición socioeconómica o conocimiento de la enfermedad. Además, cuando se da este sesgo la enfermedad tiende a diagnosticarse antes en expuestos que en no expuestos y la sobrestimación del efecto resulta en buena medida de los casos con enfermedad menos grave o avanzada, por lo que estratificación del análisis según gravedad de la enfermedad en el momento del diagnóstico puede ayudar a valorar la posible presencia de este sesgo. (8)
- Sesgo de pérdida de seguimiento: Se puede valorar la influencia de las pérdidas sobre la exposición y otras variables comparando los participantes perdidos con los que permanecen, pero es más difícil valorar su influencia sobre el desenlace (enfermedad). Se puede reducir motivando e incentivando a los participantes, y realizando un buen seguimiento y control de la recogida de información. En el momento del análisis se puede mitigar su impacto con técnicas estadísticas como la ponderación de los resultados por el inverso de la probabilidad de censura. (19)
- Sesgo de no respuesta: El efecto de la no respuesta se puede valorar si los participantes con y sin valores para una variable de interés pueden compararse con respecto a características básicas como edad, sexo, nivel de estudios, etc. El sesgo puede minimizarse haciendo varios intentos para captar a los sujetos elegidos y contando con un instrumento de recogida de información que facilite la respuesta. También puede considerarse la ponderación de los resultados por el inverso de la probabilidad de no respuesta. La disminución de la potencia estadística en los análisis multivariantes puede mitigarse creando en las variables de ajuste con alta proporción de valores ausentes una categoría específica para incluir estos valores.

- **Sesgo de colisión:** Los diagramas dirigidos no cíclicos son útiles para identificarlo. Para prevenir este sesgo hay que evitar restringir la población de estudio o ajustar en el análisis estadístico por factores de colisión (características que probablemente se vean afectadas tanto por la exposición como por la enfermedad) y conviene minimizar las pérdidas de seguimiento. (16)

2.5. Sesgos de información o de clasificación errónea

Los sesgos de información son errores sistemáticos que se producen en un estudio epidemiológico durante la recogida e interpretación de la información. Estos errores aparecen en la fase de recogida de información sobre la exposición y/o sobre el desenlace, generando una incorrecta clasificación de una parte más o menos importante de los participantes de un estudio. Es por esto que, en ocasiones, se conocen como sesgos de clasificación errónea.

Su magnitud va a depender de la sensibilidad y la especificidad del instrumento o aparato que utilicemos para medir la exposición o el desenlace.

Los sesgos de información se pueden clasificar según distintos criterios, por ejemplo, según el tipo de clasificación errónea que generan, diferencial vs. no diferencial, o según su origen.

2.5.1. Sesgos de clasificación errónea no diferencial

Los sesgos de clasificación errónea no diferencial son aquellos en los que el error al clasificar a las personas del estudio es el mismo en los grupos que queremos comparar.

Clasificación errónea no diferencial de la enfermedad

En los sesgos de clasificación errónea no diferencial de la enfermedad, el error al clasificar a las personas según la presencia o ausencia del desenlace es igual en las expuestas que en las no expuestas, es decir, es independiente de la exposición.

Veamos un ejemplo con datos ficticios:

En un estudio de cohortes, queremos analizar la asociación entre la celiaquía —intolerancia al gluten— y el sexo. Realizamos el seguimiento de 20 000 personas recién nacidas durante los tres primeros años de vida. La detección de anticuerpos antitransglutaminasa tisular 2 (anti-TG2) es la prueba más común en el despistaje de la celiaquía. Su sensibilidad para el diagnóstico es del 95 % y su especificidad del 90 %, tanto en hombres como en mujeres. (20) Esto significa que, el 95 % de las personas intolerantes al gluten y el 90 % de las personas no intolerantes van a quedar bien clasificadas mediante este método diagnóstico, independiente de que sean mujeres u hombres. Si no hubiese sesgos en la prueba diagnóstica (sensibilidad y especificidad del 100 %) los datos que obtendríamos serían estos (Tabla 4):

Tabla 4. Ejemplo con datos ficticios. Estudio de cohortes. Sensibilidad: 100 %. Especificidad: 100 %

Situación SIN sesgo	Niñas (expuestas)	Niños (no expuestos)	Total
Celiaquía	180	80	260
No celiaquía	9820	9920	19 740
Total	10 000	10 000	20 000
	IA: 0,018	IA: 0,008	RR: 2,25
	El 1,8 % de las niñas son diagnosticadas tras tres años de seguimiento.	El 0,8 % de los niños son diagnosticados tras tres años de seguimiento.	El riesgo de tener celiaquía es 2,25 veces mayor en niñas que en niños.

DR: 0,01. En los 3 primeros años de vida, el exceso de riesgo de desarrollar celiaquía en las niñas, comparado con los niños, es de 1 punto porcentual.

IA: Incidencia acumulada; RR: Razón de riesgo; DR: Diferencia de riesgos.

Pero como la prueba para diagnosticar la celiaquía (anti-TG2), tiene una sensibilidad de un 95 % y una especificidad de un 90 %, los datos que observaríamos serían estos (Tabla 5):

Tabla 5. Ejemplo con datos ficticios de clasificación errónea no diferencial de la enfermedad en un estudio de cohortes. Sensibilidad: 95 %. Especificidad: 90 %

Situación CON sesgo	Niñas (expuestas)	Niños (no expuestos)	Total
Celiaquía	1153	1068	2221
No celiaquía	8847	8932	17 779
Total	10 000	10 000	20 000
	IA: 0,12	IA: 0,11	RR: 1,09
	El 12 % de las niñas son diagnosticadas tras tres años de seguimiento.	El 11 % de los niños son diagnosticados tras tres años de seguimiento.	El riesgo de obtener un diagnóstico de celiaquía es un 9 % mayor en niñas que en niños.

DR: 0,01. En los 3 primeros años de vida, el exceso de riesgo de desarrollar celiaquía en las niñas, comparado con los niños, es de 1 punto porcentual.

IA: Incidencia acumulada; RR: Razón de riesgo; DR: Diferencia de riesgos.

Los sesgos de clasificación errónea no diferencial, cuando la exposición es dicotómica (exposición/no exposición), generalmente producen una infraestimación de la medida de asociación; es decir, un sesgo hacia el valor nulo. En el ejemplo anterior hemos pasado de una razón de riesgo (RR) sin sesgo de 2,25 a una de 1,09 con sesgo.

En los estudios de cohortes, si la clasificación errónea se produce exclusivamente en las personas que sufren el desenlace o efecto y afecta en la misma proporción a la población expuesta y no expuesta, el error de clasificación se trasladaría a la medida absoluta, la diferencia de riesgos (DR), pero no a la medida relativa, la razón de riesgo (RR).

Vamos a ilustrar esta situación con el siguiente ejemplo ficticio, en el que disponemos de una prueba diagnóstica para el diagnóstico de la celiaquía que tiene una sensibilidad del 50 % y una especificidad del 100 %, tanto para hombres como para mujeres. En esta situación obtendríamos la siguiente clasificación de las personas estudiadas (Tabla 6):

Tabla 6. Ejemplo con datos ficticios de clasificación errónea no diferencial de la enfermedad en un estudio de cohortes. Sensibilidad: 50 %. Especificidad: 100 %

Situación CON sesgo	Niñas (expuestas)	Niños (no expuestos)	Total
Celiaquía	90	40	130
No celiaquía	9910	9960	19 870
Total	10 000	10 000	20 000
	IA: 0,009	IA: 0,004	RR: 2,25
	El 0,9 % de las mujeres son diagnosticadas tras tres años de seguimiento.	El 0,4 % de los hombres son diagnosticados tras tres años de seguimiento.	El riesgo de tener celiaquía es 2,25 veces mayor en mujeres que en hombres.

DR: 0,005. En los 3 primeros años de vida, el exceso de riesgo de desarrollar celiaquía en las niñas, comparado con los niños, es de 0,5 puntos porcentuales.

IA: Incidencia acumulada; RR: Razón de riesgo; DR: Diferencia de riesgos.

En este caso, el efecto del sesgo lo observamos en la medida absoluta —DR— pero no en la medida relativa —RR—.

Clasificación errónea no diferencial de la exposición

En estos sesgos el error que se produce al clasificar a las personas según la presencia o ausencia de un determinado factor de riesgo o exposición es igual en las que presentan el desenlace que en las que no lo presentan, es decir, es independiente del desenlace. Es un sesgo que ocurre, en mayor medida, en los estudios de casos y controles.

Veamos un ejemplo ficticio:

Diseñamos un estudio de casos y controles para medir la asociación del consumo de alcohol y la depresión en jóvenes. Si el consumo de alcohol lo midiésemos sin errores (sensibilidad y especificidad del 100 %) obtendríamos los siguientes datos (Tabla 7):

Tabla 7. Ejemplo con datos ficticios. Estudio de casos y controles. Sensibilidad: 100 %. Especificidad: 100 %

Situación SIN sesgo	Exposición al alcohol	No exposición al alcohol	Total
Caso –depresión–	5400	8100	13 500
Control –no depresión–	600	5900	6500

OR (odds ratio): 6,55. Las personas expuestas al alcohol tienen una probabilidad 6,55 veces de sufrir depresión respecto a las no expuestas.

Ahora bien, el consumo de alcohol se midió con la “Prueba de Identificación de Trastornos por Uso de Alcohol —AUDIT—, cuya sensibilidad y especificidad en la población joven se estima en 86 % y 90 %, respectivamente (21), tanto en personas deprimidas como no deprimidas. Usando esta prueba, los datos que observaríamos serían estos (Tabla 8):

Tabla 8. Ejemplo con datos ficticios de clasificación errónea no diferencial de la exposición en un estudio de casos y controles. Sensibilidad: 86 %. Especificidad: 90 %

Situación CON sesgo Sensibilidad 86 % Especificidad 90 %	Exposición al alcohol	No exposición al alcohol	Total
Caso –depresión–	5454	8046	13 500
Control –no depresión–	1106	5394	6500

OR (odds ratio): 3,31. Las personas expuestas al alcohol tienen una probabilidad 3,31 veces de sufrir depresión respecto a las no expuestas.

Observamos que se produce una infraestimación de la medida de asociación, obteniendo valores que tienden a la nulidad pasando de una OR sin sesgo de 6,55 a una OR sesgada de 3,31.

2.5.2. Sesgos de clasificación errónea diferencial

En estos el error que se comete es diferente en los grupos que queremos comparar. Pueden actuar en cualquier dirección, tanto sobreestimando como infraestimando la magnitud de la asociación, por eso, es muy importante predecir tanto su presencia como su dirección para valorar en qué medida pueden afectar a los resultados del estudio.

Clasificación errónea diferencial de la enfermedad

En estos sesgos el error al clasificar a los participantes según la presencia o ausencia del desenlace es diferente en las personas expuestas que en las no expuestas; es decir, depende de la exposición. Se da sobre todo en los estudios longitudinales, como cohortes y ensayos clínicos.

Veamos un ejemplo con datos ficticios:

Realizamos un estudio de cohortes en mayores de 60 años para analizar la asociación del sexo con la enfermedad coronaria obstructiva. Tras 10 años de seguimiento, si usáramos un método diagnóstico con una sensibilidad y una especificidad del 100 %, los resultados que obtendríamos serían estos (Tabla 9):

Tabla 9. Ejemplo con datos ficticios. Estudio de cohortes. Sensibilidad: 100 %. Especificidad: 100 %

Situación SIN sesgo	Hombres (expuestos)	Mujeres (no expuestas)	Total
Enfermedad coronaria obstructiva	3100	2400	5500
No enfermedad coronaria obstructiva	96 900	97 600	194 500
Total	100 000	100 000	200 000
	IA: 3,1 %	IA: 2,4 %	RR: 1,29
	El 3,1 % de los hombres sufren enfermedad coronaria obstructiva tras 10 años de seguimiento.	El 2,4 % de las mujeres sufren enfermedad coronaria obstructiva tras 10 años de seguimiento.	El riesgo de enfermedad coronaria obstructiva es un 29 % mayor en hombres que en mujeres.

DR:0,007 Comparados con las mujeres en los hombres se produce un exceso de riesgo de desarrollar enfermedad coronaria de 7 casos por cada mil personas seguidas 10 años.

IA: Incidencia acumulada; RR: Razón de riesgo; DR: Diferencia de riesgo.

Sin embargo, el diagnóstico de enfermedad coronaria obstructiva se realizó mediante una angio-TAC coronaria con contraste que tiene en mujeres una sensibilidad del 90 % y una especificidad del 96 % y en hombres una sensibilidad de 89 % y una especificidad del 78 %. (22)

Con el uso de dicha técnica, los datos que observaríamos serían estos (Tabla 10):

Tabla 10. Ejemplo con datos ficticios de clasificación errónea diferencial de la enfermedad en un estudio de cohortes. Mujeres: sensibilidad 90 %, especificidad 96 %; Hombres: sensibilidad 89 %, especificidad 78 %

Situación CON sesgo	Hombres	Mujeres	Total
Enfermedad coronaria obstructiva	24 077	6064	30 141
No enfermedad coronaria obstructiva	75 923	93 936	169 859
Total	100 000	100 000	200 000
	IA: 24,1 %	IA: 6,1 %	RR: 4,0
	El 24,1 % de los hombres sufren enfermedad coronaria obstructiva tras 10 años de seguimiento.	El 6,1 % de las mujeres sufren enfermedad coronaria obstructiva tras 10 años de seguimiento.	El riesgo de enfermedad coronaria obstructiva es 4 veces mayor en hombres que en mujeres.

DR: 0,18. Comparados con las mujeres en los hombres se produce un exceso de riesgo de desarrollar enfermedad coronaria de 180 casos por cada mil personas seguidas 10 años.

IA: Incidencia acumulada; RR: Razón de riesgo; DR: Riesgo atribuible en expuestos.

El uso de angio-TAC coronaria con contraste para el diagnóstico de la enfermedad coronaria obstructiva produjo una sobreestimación de la asociación del sexo con dicha enfermedad.

Clasificación errónea diferencial de la exposición

Ocurre cuando el error que se comente al clasificar a las personas según su exposición, es diferente en las personas en las que ocurre el desenlace que en las que no ocurre; es decir, el error depende del desenlace.

Veamos un ejemplo con datos ficticios:

Queremos identificar la asociación entre uso de servicios sanitarios (atención hospitalaria/atención primaria) y consumo problemático de alcohol. Para ello, ponemos en marcha un estudio de casos y controles en el que los casos fueron pacientes hospitalarios y los controles pacientes de

atención primaria. La información sobre consumo de alcohol la recogemos mediante el cuestionario CAGE (Cut-down, Annoyed, Guilty, Eye-opener questionnaire). Si en la recogida de información no hubiese sesgos (sensibilidad 100 %; especificidad 100 %), los datos que obtendríamos serían estos (Tabla 11):

Tabla 11. Ejemplo con datos ficticios. Estudio de casos y controles. Sensibilidad: 100 %. Especificidad: 100 %

Situación SIN sesgo	Expuesto alcohol	No expuesto alcohol	Total
Caso (paciente hospitalario)	1700	8300	10 000
Control (paciente atención primaria)	1400	8600	10 000

OR (odds ratio): 1,26. La probabilidad de acudir a hospital (necesitar atención hospitalaria) en los expuestos a alcohol es 1,26 veces la probabilidad en los no expuestos a alcohol.

Ahora bien, el cuestionario CAGE ha mostrado una sensibilidad del 76 % y una especificidad del 94 % cuando se utiliza a nivel hospitalario, mientras que, en atención primaria, la sensibilidad es del 91 % y la especificidad del 84 %. (23) Por tanto, en estas condiciones los datos que obtendríamos serían estos (Tabla 12):

Tabla 12. Ejemplo con datos ficticios de clasificación errónea diferencial de la exposición en un estudio de casos y controles. Nivel hospitalario: sensibilidad 76 %, especificidad 94 %; Atención primaria: sensibilidad 91 %, especificidad 84 %

Situación CON sesgo	Expuesto alcohol	No expuesto alcohol	Total
Caso (paciente hospitalario)	1790	8210	10 000
Control (paciente atención primaria)	2650	7350	10 000

OR (odds ratio): 0,6 La probabilidad de acudir a hospital (necesitar atención hospitalaria) en los expuestos a alcohol es un 40 % menor que en los no expuestos a alcohol.

Observamos que se ha invertido el sentido de la asociación.

2.5.3. Sesgos de información según su origen

Sesgo de migración diagnóstica

Suele aparecer en los estudios cohortes de larga duración cuando se producen cambios en la medición de la exposición o en el diagnóstico del desenlace durante el seguimiento. Estos cambios pueden deberse a modificaciones en los criterios

diagnósticos, cambios en los umbrales de las exposiciones, mejoras en los instrumentos de medida o, en el caso de desenlaces autoreferidos, a cambios en la actitud, percepción o conocimiento de la población.

Por ejemplo, imaginemos una situación ficticia, donde queremos realizar un estudio retrospectivo para analizar la supervivencia de pacientes oncológicas tras 20 años de seguimiento según la gravedad de las lesiones oncológicas en el momento del diagnóstico.

Durante este periodo es posible que las técnicas de diagnóstico hayan evolucionado y que en el momento actual nos permitan identificar con mayor precisión la gravedad de la lesión. Por tanto, pacientes que al inicio del estudio fueron clasificados como pacientes en estadios iniciales de la enfermedad, es probable que actualmente hubiesen sido clasificados en estadios más avanzados porque los instrumentos actuales son más precisos. Por tanto, la supervivencia actual de los pacientes oncológicos graves, es posible que aumente al incorporar pacientes de menor gravedad. De la misma manera, la supervivencia del grupo de pacientes con gravedad baja aumentará porque los pacientes de este grupo con lesiones de mayor gravedad habrán sido clasificados en un estadio superior.

De forma análoga al ejemplo anterior, en los estudios de cohortes de larga duración se pueden dar cambios o “migraciones” en la medición de la exposición.

Sesgo de sospecha diagnóstica

El sesgo de sospecha diagnóstica, tal y como hemos visto anteriormente, es un sesgo de selección característico de los estudios de casos y controles. En los estudios de cohortes, el sesgo de sospecha diagnóstica es un sesgo de información. Se produce cuando, una vez seleccionadas a las personas del estudio, se hace mayor esfuerzo en diagnosticar la enfermedad si se sabe que la persona está expuesta al factor de riesgo. Es un sesgo de clasificación errónea diferencial de la enfermedad que tiende a sobreestimar la medida de asociación.

Por ejemplo, si una persona procedente de un país de Latinoamérica, donde la prevalencia de tuberculosis es alta, acude a la consulta de atención primaria con tos persistente la probabilidad de que se prescriba la prueba de Mantoux por sospecha de tuberculosis es mayor que si la persona ha nacido en España. En población latinoamericana, la mayor sospecha dará lugar a una sobreestimación de la tuberculosis, respecto a la población española.

Sesgo de inaceptabilidad

Es un sesgo de clasificación errónea de la exposición, generalmente de tipo diferencial que produce una infraestimación de la asociación, aunque en determinados estudios puede ser no diferencial. Es más frecuente en los estudios de casos

y controles, cuando las personas no reconocen una exposición porque socialmente no es bien aceptada.

Por ejemplo, si queremos analizar el efecto del consumo de alcohol en el embarazo sobre el bajo peso al nacer, es posible que las madres que hayan dado a luz niños o niñas con bajo peso reconozcan menos el consumo de alcohol debido al sentimiento de culpa. En este caso, estaríamos ante una clasificación errónea diferencial de la exposición, que producirá una infraestimación del efecto del alcohol en el bajo peso al nacer.

Sesgo de obsequiosidad

Es un sesgo similar al anterior, pero en sentido opuesto. Generalmente produce clasificación errónea en la exposición, generalmente diferencial, aunque puede ser no diferencial. Aparece cuando las personas sobrevaloran conductas que están bien aceptadas socialmente, por ejemplo, cuando se recoge información sobre la exposición a comportamientos saludables como consumo de frutas y verduras o ejercicio físico, etc.

Por ejemplo, si queremos analizar la asociación entre el tiempo dedicado a las tareas del hogar y las migrañas, es posible que los hombres, que generalmente realizan menos trabajo en el hogar, sobreestimen el tiempo que dedican a dichas tareas porque es un comportamiento aceptado socialmente. Esto va a producir una infraestimación en la asociación entre el tiempo dedicado a las tareas del hogar y las migrañas.

Sesgo de atención o efecto Hawthorne

Es un sesgo que puede darse en estudios longitudinales. Consiste en que las personas pueden cambiar su comportamiento habitual por formar parte de un estudio o, en general, por saberse observados, cambios que pueden atribuirse erróneamente a la exposición o intervención estudiada. Generalmente produce una clasificación errónea no diferencial de la exposición, aunque en ocasiones puede ser diferencial. Debe su nombre a que fue descrito en Hawthorne, Nueva Jersey, donde la Western Electric tenía una fábrica. Allí se hizo un experimento para ver cómo los cambios en la iluminación, humedad o mobiliario, afectaban a la productividad laboral. Observaron que cualquier cambio producía una mejora de la productividad. (24)

Sesgo de recuerdo

Es un sesgo de clasificación errónea de la exposición. Generalmente es un sesgo diferencial, pero en algunos estudios podría ser no diferencial. Es frecuente en los estudios de casos y controles, donde la información se suele recoger retrospectivamente. En estas situaciones, si la exposición ocurrió hace mucho tiempo,

es posible las personas no tengan un recuerdo exacto de la misma. Dependiendo de si el olvido es mayor en el grupo expuesto, en el no expuesto o es similar en ambos, se va a producir infraestimación, sobreestimación o ningún efecto sobre la asociación.

Por ejemplo, las madres que tuvieron hijos con malformaciones congénitas podrían recordar mejor el consumo durante el embarazo de fármacos como la talidomida que las que no tuvieron hijos con estos problemas, produciéndose una sobreestimación de la exposición sobre el efecto. También es posible que, debido a las consecuencias traumáticas del evento, las mujeres que fueron tratadas con talidomida y tuvieron un hijo con malformaciones, hayan olvidado inconscientemente el consumo del fármaco, con lo que estaríamos ante una infraestimación de la exposición. En ambos supuestos de este ejemplo estamos ante una clasificación errónea diferencial de la exposición.

Sesgo del entrevistador/observador

Es un sesgo de clasificación errónea diferencial de la exposición. En los estudios de casos y controles, cuando la persona que recoge la información sabe que está ante un caso, es posible que haga un mayor esfuerzo en identificar la exposición que cuando está ante un control. En esta situación la clasificación errónea diferencial de la exposición produciría sobreestimación de la asociación.

Sesgo de latencia de la enfermedad

Es un sesgo de clasificación errónea diferencial de la exposición que generalmente produce infraestimación de la asociación. Se debe a que a veces la sintomatología de la enfermedad produce un cambio en la exposición antes de que se diagnostique.

Por ejemplo: una persona con molestias gástricas por una úlcera de estómago no diagnosticada es posible que no tome salicilatos porque observa un agravamiento de la dolencia al consumirlos. Si el entrevistador, en el momento del diagnóstico, recoge información sobre el consumo reciente de salicilatos, este fármaco puede ser identificado como un factor protector de la úlcera gástrica cuando realmente es un factor de riesgo.

Sesgo de desenmascaramiento

Es un sesgo de clasificación errónea diferencial de la enfermedad que generalmente produce una sobreestimación de la asociación. Se da principalmente en los estudios de cohortes, cuando el conocimiento de la exposición aumenta la sospecha de la enfermedad.

Por ejemplo, se quiere analizar la asociación del consumo de estrógenos con el cáncer de endometrio. Tanto la administración de estrógenos como el cáncer de endometrio producen metrorragia, por tanto, es posible que se busque con más intensidad el cáncer de endometrio en las mujeres que consumen estrógenos.

2.6. Control de los sesgos de información

Si bien los sesgos de información se dan en todos los estudios epidemiológicos, hay que usar todas las herramientas disponibles para prevenirlos, como:

- Estandarizar los procedimientos de recogida de información, de forma que sean constantes en el espacio, tiempo, contexto, grupos y personas, incluyendo para expuestos y no expuestos, y enfermos y no enfermos. Esto implica:
 - ♦ Desarrollar protocolos detallados y claros sobre el trabajo de campo.
 - ♦ Calibrar y validar los instrumentos de medida o recogida de información, utilizando, cuando sea posible, instrumentos con sensibilidad y especificidad conocidas y elevadas.
 - ♦ No cambiar dichos instrumentos a lo largo del tiempo y, si se hace, comprobar que los cambios no afectan a los resultados con entrevista cognitiva y pilotaje.
 - ♦ Minimizar el número de personas que participan en la recogida de la información y formarlas de forma homogénea, enmascarar, cuando sea posible, su conocimiento sobre quién está expuesto o no expuesto, enfermo o no enfermo, y prohibir los juicios de valor sobre las conductas o actitudes de los participantes en el estudio.
- La flexibilización de los procedimientos de recogida de información puede disminuir la no-respuesta, pero comporta un riesgo evidente de sesgos de información.
- Reducir al mínimo el tiempo entre la exposición y la recogida de información para mejorar la homogeneidad de los grupos que se van a comparar.
- Incorporar, cuando el diseño lo permita, un grupo control, idealmente con placebo, que minimice el sesgo de atención.

2.7. Otros sesgos

Regresión a la media

En ocasiones se clasifica a las personas en expuestas/no expuestas, o en enfermas/no enfermas cuando sobrepasan un umbral en un parámetro biológico. Por ejemplo, las personas que tienen un colesterol total en sangre de 240 mg/dl o mayor se considera que tienen un colesterol alto. Se diagnostica diabetes si el nivel de glucosa en ayunas es de 126 mg/dl o mayor. Cuando clasificamos a las personas a partir de una medición y obtenemos valores extremos, es posible que, si

realizamos sucesivas mediciones, estos valores extremos se acerquen a la media, es decir, estén “regresando a la media”. Para que se dé hay que clasificar a las unidades o personas analizadas basándose en un parámetro cuantitativo con variabilidad intraindividual. Esta variabilidad puede ser fisiológica o porque midamos un parámetro sensible a cambios ambientales, o por errores aleatorios en los instrumentos de medida.

Un ejemplo de regresión a la media es la llamada “hipertensión de bata blanca”. La medición de la presión arterial, es un proceso sencillo, habitual en la práctica clínica y que permite diagnosticar la hipertensión esencial. Sin embargo, la medición de la hipertensión puede mostrar una gran variabilidad, entre otras cosas, por la reacción de alerta que experimentan algunas personas ante la presencia médica. Esta situación de alerta puede hacer que la presión arterial de una persona en el contexto clínico, sea superior a la que tiene en su casa cuando la medición la realiza un familiar. Si en una primera consulta médica se obtiene una tensión arterial alta, es posible que, en sucesivas mediciones en la misma consulta, los valores vayan regresando a la media.

Cuando la población de estudio presenta un valor extremo de un parámetro biológico, conviene utilizar para cada registro individual la media de varias mediciones, al objeto de minimizar el efecto de regresión a la media en la recogida de información.

Falacia ecológica

Es un sesgo que se produce en los estudios ecológicos, donde las unidades de análisis no son individuos sino grupos de personas o de datos agregados. El error se produce cuando extrapolamos al nivel individual los resultados obtenidos a nivel grupal o poblacional. Este sesgo queda expuesto a través de los hallazgos de Emile Durkheim, cuando analizó las tasas de suicidios en las provincias de Prusia entre 1883 y 1890. (25) El sociólogo francés encontró una asociación entre la proporción de protestantes en cada provincia y sus tasas de suicidio. De este trabajo no puede extraerse la conclusión de que las personas protestantes se suicidan más que las católicas, ya que desconocemos quién se suicidaba, si las personas protestantes o las católicas que vivían en regiones de mayoría protestante. Debido a la falacia ecológica la extrapolación de los resultados obtenidos a nivel grupal al nivel individual debe hacerse con mucha prudencia.

Fallos en el registro de datos

En ocasiones, se producen fallos en los registros de la información que pueden distorsionar nuestros resultados. Estos pueden ser errores humanos (en la codificación de los datos, en la extracción de información de los registros, en la calibración de los instrumentos de medida etc.) o errores asociados a los instrumentos de medida, así como a la variabilidad de los mismos.

Sesgo de causalidad inversa

Se produce cuando se interpreta como causal una asociación sin evidencias suficientes para asegurar que la exposición o supuesta causa precede en el tiempo a la enfermedad. A veces se denomina sesgo temporal. Se da sobre todo en estudios transversales o de series temporales en los que los datos de exposición y enfermedad se recogen en el mismo momento o período y por tanto falta información sobre la secuencia temporal entre la exposición y la enfermedad. Sin embargo, también puede aparecer en estudios de casos y controles cuando la exposición se mide después del diagnóstico de enfermedad, o incluso en estudios de cohortes con desenlace de mortalidad. Este sesgo podría, por ejemplo, explicar la asociación negativa encontrada en estudios transversales en niños entre poseer una mascota y alergia/asma, ya que las familias tienden a desprenderse o no adquirir mascotas cuando aparecen estos problemas o la mayor mortalidad a corto plazo por cáncer de pulmón en exfumadores que en fumadores actuales, ya que algunos fumadores tienden a abandonar el hábito cuando aparecen síntomas o son diagnosticados de enfermedades relacionadas con el tabaco. (8)

Sesgos al seleccionar estudios para sintetizar resultados

Pueden sesgar las conclusiones procedentes de revisiones de la literatura, particularmente de revisiones sistemáticas y metanálisis. El más conocido es el sesgo de publicación, pero hay otros.

- **Sesgo de publicación:** Se puede dar cuando los informes publicados no representan los estudios insesgados realizados sobre la estimación de un parámetro de interés. En buena medida se debe a que ante estudios de igual calidad es más probable que se publiquen los que obtienen resultados positivos (efecto estadísticamente significativo con rechazo de hipótesis nula) que negativos o inconcluyentes por insuficiente tamaño muestral. Esto produce una perspectiva incompleta y sesgada de la evidencia disponible muy preocupante, máxime cuando las “expediciones de pesca” sin hipótesis previas sólidas, que implican un análisis intensivo de datos, pueden encontrar de vez en cuando por mero efecto del azar efectos positivos que no existen en la población objetivo. Además de la significación estadística, hay otros factores espurios que pueden generar sesgo de publicación como la financiación o prestigio de los autores o las instituciones que los avalan. (8,9,12) Para minimizarlo el Comité Internacional de Directores de Revistas Médicas recomienda que los editores no deben excluir de consideración para publicación los estudios con hallazgos no estadísticamente significativos (negativos) o no concluyentes, ya que dichos estudios, combinados con la de otros similares mediante metanálisis, pueden proporcionar evi-

dencia para responder preguntas importantes. Además, un registro público de esos hallazgos puede impedir la replicación injustificada del esfuerzo o ser valioso para otros investigadores que estén considerando un trabajo similar. Otra estrategia para reducirlo es la publicación previa de diseños de investigación. (8,26)

- Sesgos de selección de estudios publicados: La probabilidad seleccionar un estudio publicado para revisión sistemática, metanálisis u otro fin es mayor para los más citados, publicados en inglés o en revistas de alto factor de impacto, lo que a veces podría dar lugar al análisis de muestras no representativas y sesgar las conclusiones. (8,9)

3. Referencias

1. Schwartz S, Campbell UB, Gatto NM, *et al.* Toward a clarification of the taxonomy of “bias” in epidemiology textbooks. *Epidemiology* 2015; 26(2):216-22. doi: [10.1097/EDE.0000000000000224](https://doi.org/10.1097/EDE.0000000000000224).
2. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. *Br Med J* 1956;2(5001):1071-81. doi: [10.1136/bmj.2.5001.1071](https://doi.org/10.1136/bmj.2.5001.1071).
3. Guyatt GH, Ebrahim S, Alonso-Coello P, *et al.* GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. *J Clin Epidemiol* 2017;87:14-22. doi: [10.1016/j.jclinepi.2017.05.005](https://doi.org/10.1016/j.jclinepi.2017.05.005).
4. Viswanathan M, Patnode CD, Berkman ND, *et al.* Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. *J Clin Epidemiol* 2018;97:26-34. doi: [10.1016/j.jclinepi.2017.12.004](https://doi.org/10.1016/j.jclinepi.2017.12.004).
5. Steenland K, Schubauer-Berigan MK, Vermeulen R, *et al.* Risk of Bias Assessments and Evidence Syntheses for Observational Epidemiologic Studies of Environmental and Occupational Exposures: Strengths and Limitations. *Environ Health Perspect* 2020;128(9):95002. doi: [10.1289/EHP6980](https://doi.org/10.1289/EHP6980).
6. Lash TL, Fox MP, MacLehose RF, *et al.* Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43(6):1969-85. doi: [10.1093/ije/dyu149](https://doi.org/10.1093/ije/dyu149).

7. Lash TL, VanderWeele TJ, Haneuse S, *et al.* Modern epidemiology. 4th edition. Philadelphia, PA: Wolters Kluwers 2021.
8. Szklo M, Nieto FJ. Epidemiology beyond the basics, 4th edition. Burlington, MA: Jones & Bartlett Learning 2019.
9. Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;58(8):635-41. doi: [10.1136/jech.2003.008466](https://doi.org/10.1136/jech.2003.008466).
10. Caldwell GG, Kelley DB, Heath CW, Jr. Leukemia among participants in military maneuvers at a nuclear bomb test. A preliminary report. *JAMA* 1980;244(14):1575-8.
11. Sadetzki S, Bensal D, Novikov I, *et al.* The limitations of using hospital controls in cancer etiology--one more example for Berkson's bias. *Eur J Epidemiol* 2003;18(12):1127-31. doi: [10.1023/b:ejep.00000006634.49205.c5](https://doi.org/10.1023/b:ejep.00000006634.49205.c5).
12. De Irala J, Martínez MA, Seguí M. Epidemiología aplicada. 2.^a Edición. Barcelona: Ariel España 2017.
13. Celentano DD, Szklo M. Gordis. Epidemiology. 6th edition. Philadelphia, PA: Elsevier 2019.
14. Williams GM, Najman JM, Clavarino A. Correcting for numerator/denominator bias when assessing changing inequalities in occupational class mortality, Australia 1981 -2002. *Bull World Health Organ* 2006;84(3):198-203. doi: [10.2471/blt.05.028894](https://doi.org/10.2471/blt.05.028894).
15. Greenland S. Response and follow-up bias in cohort studies. *Am J Epidemiol* 1977;106(3):184-7. doi: [10.1093/oxfordjournals.aje.a112451](https://doi.org/10.1093/oxfordjournals.aje.a112451).
16. Holmberg MJ, Andersen LW. Collider Bias. *JAMA* 2022;327(13):1282-83. doi: [10.1001/jama.2022.1820](https://doi.org/10.1001/jama.2022.1820).
17. Accorsi EK, Qiu X, Rumpler E, *et al.* How to detect and reduce potential sources of biases in studies of SARS-CoV-2 and COVID-19. *Eur J Epidemiol* 2021;36(2):179-96. doi: [10.1007/s10654-021-00727-7](https://doi.org/10.1007/s10654-021-00727-7).
18. Schwartzbaum J, Ahlbom A, Feychting M. Berkson's bias reviewed. *Eur J Epidemiol* 2003;18(12):1109-12. doi: [10.1023/b:ejep.00000006552.89605.c8](https://doi.org/10.1023/b:ejep.00000006552.89605.c8).
19. Howe CJ, Cole SR, Lau B, *et al.* Selection Bias Due to Loss to Follow Up in Cohort Studies. *Epidemiology* 2016;27(1):91-7. doi: [10.1097/EDE.0000000000000409](https://doi.org/10.1097/EDE.0000000000000409).
20. Fabiano F, Lista D, Torres J, *et al.* Sensibilidad y especificidad de los marcadores serológicos para diagnóstico y seguimiento de la condición celíaca. *Gen* 2014;68(1):8-11.

21. Rial A, Golpe S, Araujo M, *et al.* Validación del “Test de Identificación de Trastornos por Consumo de Alcohol” (AUDIT) en población adolescente española. *Psicología Conductual* 2017;25(2):371-86.
22. Acevedo M, Varleta P. Exámenes diagnósticos cardiovasculares: ¿diferencias de género en su interpretación? *Revista Colombiana de Cardiología*; 2018(25):66-72.
23. Sáiz PA, González MP, Paredes MB, *et al.* Instrumentos de evaluación en alcoholismo. *Adicciones* 2002;14(5):387-403.
24. Franke RH, Kaul JD. The Hawthorne experiments: First statistical interpretation. *American Sociological Review* 1978:623-43.
25. Durkheim E. Suicide. A study in sociology. London: Routledge & K. Paul 1952.
26. ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. Updated May 2022. 2022. <https://www.icmje.org/icmje-recommendations.pdf>.

Capítulo 5. Confusión e interacción

BELÉN SANZ BARBERO

1. Confusion

- 1.1. Concepto de confusión
- 1.2. Condiciones que debe cumplir una variable para ser considerada un factor de confusión
- 1.3. Efecto de la confusión sobre la medida de asociación observada
- 1.4. Cuantificación de la confusión
- 1.5. Control de la confusión
- 1.6. La confusión residual

2. Interacción o modificación del efecto

3. Referencias

1. Confusion

1.1. Concepto de confusión

La confusión es un sesgo que produce una distorsión de la asociación entre una exposición y un efecto debido a la presencia de otras variables, que llamamos factores de confusión, confusores o variables confusoras.

La confusión es resultado de las relaciones existentes en las variables de una población. A diferencia de los sesgos de selección y de información, no es posible evitarla, sin embargo, cuando la conocemos, podemos corregirla o controlarla. (1,2)

Para ejemplificar la confusión, vamos a considerar los resultados de un estudio hipotético de cohortes que tiene como objetivo estimar la asociación de la exposición “trabajar en exteriores” en el desarrollo de una enfermedad respiratoria. Tras 10 años de seguimiento observamos que, en las personas que trabajan en exteriores se produjeron 200 casos nuevos de enfermedad por cada 990 personas, mientras que en las personas que trabajan en interiores se produjeron 75 casos nuevos de enfermedad por cada 1025 personas (tabla 1). Por tanto, podemos decir que las personas que trabajan exteriores multiplican por 2,76 el riesgo de contraer la enfermedad que tienen las que trabajan en interiores.

Tabla 1. Estudio hipotético. Distribución de la enfermedad respiratoria según el trabajo en interiores/exteriores

	Trabajo exteriores	Trabajo interiores	Total
Enfermedad	200	75	275
No enfermedad	790	950	1740
Total	990	1025	2015
Incidencia acumulada	0,20	0,07	0,14
	En las personas que trabajan en exteriores se producen 20 casos de enfermedad por cada 100 personas seguidas durante 10 años.	En las personas que trabajan en interiores se producen 7 casos de enfermedad por cada 100 personas seguidas durante 10 años.	En la cohorte analizada se producen 14 casos de enfermedad por cada 100 personas seguidas durante 10 años.
Razón de Riesgos (RR)	RR: 2,76. Las personas que trabajan en exteriores multiplican por 2,76 el riesgo de contraer la enfermedad que tienen las que trabajan en interiores		

Hasta ahora, hemos considerado dos variables: la exposición “trabajar en el exterior” y el efecto, que es la enfermedad. Ahora bien, podríamos preguntarnos si hay más fumadores entre las personas que trabajaban en exteriores que entre las que trabajaban en interiores. De ser así, este hecho podría explicar, al menos en parte, por qué las personas que trabajaban en el exterior tienen un riesgo mayor de desarrollar la enfermedad respiratoria que las que trabajaban en interiores.

Para responder a esta pregunta, dividimos a los participantes de nuestro estudio en dos estratos en función del consumo de tabaco: personas fumadoras (tabla 2) y no fumadoras (tabla 3). En ninguno de los estratos analizados existe una asociación clara entre trabajar en exterior y la enfermedad (RR fumadoras: 1,03; RR no fumadoras: 1,1). Estos resultados sugieren que la asociación entre trabajar en exteriores y la enfermedad respiratoria es muy débil si se tiene en cuenta el efecto del tabaco. Por tanto, decimos que el tabaco es un factor de confusión.

Tabla 2. Estudio hipotético. Distribución de la enfermedad respiratoria según el trabajo en interiores/exteriores. Estrato de personas fumadoras

ESTRATO DE PERSONAS FUMADORAS			
	Trabajo exteriores	Trabajo interiores	Total
Enfermedad	194	21	215
No enfermedad	706	79	785
Total	900	100	1000
Incidencia acumulada	0,22	0,21	0,22
	En las personas FUMADORAS que trabajan en exteriores se producen 22 casos de enfermedad por cada 100 personas seguidas durante 10 años.	En las personas FUMADORAS que trabajan en interiores se producen 21 casos de enfermedad por cada 100 personas seguidas durante 10 años.	En las personas FUMADORAS se producen 22 casos de enfermedad por cada 100 personas seguidas durante 10 años.
Razón de Riesgos (RR):	RR: 1,03. Entre las personas fumadoras, aquellas que trabajan en exteriores tienen un 3 % más de riesgo de contraer la enfermedad que las que trabajan en interiores.		

Tabla 3. Estudio hipotético. Distribución de la enfermedad respiratoria según el trabajo en interiores/exteriores. Estrato de personas no fumadoras

ESTRATO DE PERSONAS NO FUMADORAS			
	Trabajo exteriores	Trabajo interiores	Total
Enfermedad	6	54	60
No enfermedad	84	871	955
Total	90	925	1015
	0,07	0,06	0,06
Incidencia acumulada	En las personas NO FUMADORAS que trabajan en exteriores se producen 7 casos de enfermedad por cada 100 personas seguidas durante 10 años.	En las personas NO FUMADORAS que trabajan en interiores se producen 6 casos de enfermedad por cada 100 personas seguidas durante 10 años.	En las personas NO FUMADORAS se producen 6 casos de enfermedad por cada 100 personas seguidas durante 10 años.
Razón de Riesgos (RR)	RR: 1,1. Entre las personas fumadoras, aquellas que trabajan en exteriores tienen un 10 % más de riesgo de contraer la enfermedad que las que trabajan en interiores.		

1.2. Condiciones que debe cumplir una variable para ser considerada un factor de confusión

Para que una variable sea un factor de confusión debe cumplir tres condiciones: (3)

1. Debe estar asociada con la exposición.
2. Debe estar asociado con la enfermedad, bien como causa o como correlato de una causa de la enfermedad.
3. No debe ser un eslabón de la cadena causal entre la exposición y la enfermedad. Dicho de otro modo, no debe ser un efecto de la exposición.

1. El factor de confusión debe estar asociado con la exposición

El factor de confusión debe distribuirse de forma diferente en las personas expuestas y no expuestas. Esta diferencia no debe ser secundaria a la asociación del factor de confusión con la enfermedad, por tanto, la asociación debe estar presente también en ausencia de enfermedad/efecto.

En el ejemplo anterior, nos preguntamos: ¿hay más fumadores entre las personas que trabajan en el exterior que entre las que trabajan en interiores?, ¿esta diferencia se observa también en las personas sanas? En nuestro estudio, el 90 % de los fumadores y el 8,9 % de los no fumadores trabajan en el exterior. Entre las personas no enfermas, el 89,9 % de las fumadoras y el 8,8 % de las no fumadoras trabajan en exteriores. La distribución del factor de confusión es distinta en personas expuestas (trabaja en el exterior) y no expuestas (trabaja en el interior). Se cumple la primera condición.

2. En factor de confusión debe estar asociado con la enfermedad

El factor de confusión debe estar asociado a la enfermedad independientemente de la exposición. Esta asociación no debe ser secundaria a la asociación del factor de confusión con la exposición, sino causal. Por tanto, la asociación entre el factor de confusión y el efecto también debe observarse en la población no expuesta.

En el ejemplo anterior, el riesgo de ser diagnosticada de enfermedad respiratoria es 3,6 veces en las personas que fuman respecto a las personas que no fuman. En el caso de las personas que trabajan en interiores (no expuestas), el riesgo de ser diagnosticadas de enfermedad respiratoria es 3,6 veces en las personas que fuman respecto a las personas que no fuman. Se cumple la segunda condición.

3. No debe ser un eslabón de la cadena causal entre la exposición y la enfermedad

Debemos preguntarnos, ¿fumar es un mediador en la asociación trabajo en exteriores-enfermedad respiratoria? No, fumar no forma parte del camino causal y por tanto cumple la tercera condición.

Podemos afirmar que la variable fumar cumple las condiciones para ser considerada a priori un factor de confusión en la asociación del trabajo en exteriores con la enfermedad respiratoria.

1.3. Efecto de la confusión sobre la medida de asociación observada

La confusión es un sesgo que va a distorsionar la medida de asociación entre la exposición y el desenlace. Esta distorsión puede hacer que observemos una asociación cuando realmente no existe relación causal, que observemos una asociación de distinta magnitud o dirección que la que observaríamos en ausencia del factor de confusión o que no observemos una asociación que realmente existe.

Cuando la confusión hace que la medida observada (confundida) se acerque al valor nulo, decimos que es una confusión negativa. La confusión negativa, por tanto, infraestima la magnitud de la asociación entre la exposición y el efecto. (3,4)

Cuando la confusión hace que la medida observada (confundida) se aleje del valor nulo, decimos que es una confusión positiva. Por tanto, la confusión positiva sobreestima la magnitud de la asociación entre la exposición y el efecto. (3,4)

Cuando la confusión hace que la medida de asociación cambie de dirección, es decir que un factor de que incrementa el riesgo lo identifiquemos como un factor protector o viceversa, decimos que la confusión es cualitativa. (3,4)

Tabla 4. Ejemplo de tipos de efecto de la confusión

Medida de asociación observada	Medida de asociación verdadera	Tipo de confusión
3,5	2,1	Positiva
0,3	0,7	Positiva
1	3,2	Negativa
1,5	3,2	Negativa
0,8	0,2	Negativa
2,0	0,7	Cualitativa
0,6	1,8	Cualitativa

1.4. Cuantificación de la confusión

El sesgo de confusión no solo podemos identificarlo, sino que también podemos cuantificarlo y controlarlo o corregirlo.

La cuantificación de la confusión se hace mediante la comparación de la medida cruda, es decir, aquella que no tiene en cuenta la confusión, con la medida ajustada. La medida ajustada es aquella que obtenemos una vez eliminado o controlado el efecto de las variables confusoras.

La comparación de la medida de asociación cruda y ajustada no se realiza mediante ningún test estadístico, sino que realizamos una comparación directa entre la magnitud de ambas medidas. La confusión se produce cuando estas medidas difieren en una magnitud relevante para el desenlace analizado. De forma orientativa, se suele considerar relevante una diferencia de al menos el 10%. (2,3,5)

Cuando la medida ajustada es mayor que la medida cruda, la cuantificación de la magnitud de la diferencia en términos porcentuales la haríamos de la siguiente manera: (2,5)

$$(\% \text{ cambio}) = \left(\frac{\text{Valor ajustado}}{\text{Valor crudo}} - 1 \right) * 100$$

Por ejemplo, si $OR_c=1$ y $OR_a=1,65$, el cambio de magnitud entre estas dos medidas es del 65 %

$$(\% \text{ cambio}) = \left(\frac{1,65}{1} - 1 \right) * 100 = 65\%$$

Cuando la medida ajustada es menor que la medida cruda la cuantificación la haríamos de la siguiente manera:

$$(\% \text{ cambio}) = 1 - \left(\frac{\text{Valor ajustado}}{\text{Valor crudo}} \right) * 100$$

Por ejemplo, si $OR_c: 1,9$ y $OR_a: 1,2$, el cambio de magnitud entre estas dos medidas es del 37 %.

$$(\% \text{ cambio}) = 1 - \left(\frac{1,2}{1,9} \right) * 100 = 37\%$$

1.5. Control de la confusión

Existen distintos procedimientos para prevenir la confusión durante la fase de diseño, así como estrategias para controlar la confusión en la fase de análisis.

1.5.1. Control de la confusión en la fase de diseño

Existen tres estrategias para prevenir o disminuir el efecto de un factor de confusión conocido durante la fase de diseño. Estas estrategias son: la aleatorización, el emparejamiento y la restricción y tienen como objetivo distribuir los factores de confusión de forma similar entre los grupos que queremos comparar, es decir entre la población expuesta y la no expuesta a un determinado factor de riesgo. Estas estrategias afectan a la forma en que seleccionamos a las personas que forman parte de nuestro estudio, por tanto, una vez que las incorporamos en nuestro diseño, no podemos revertir el cambio. (3)

a) Aleatorización

La aleatorización consiste en asignar a los y las participantes en un estudio de forma aleatoria a uno de los grupos que queremos comparar, de manera que

tengan una probabilidad predeterminada de formar parte de cada grupo. De esta forma, los posibles factores de confusión tenderán a distribuirse equitativamente entre los grupos (intercambiabilidad de los grupos de intervención). La aleatorización solo puede realizarse en los estudios experimentales. La aleatorización se rige por la ley de los grandes números y por tanto será más efectiva para conseguir la intercambiabilidad de los grupos cuanto mayor sea el número de participantes.

b) Emparejamiento

El emparejamiento se utiliza principalmente en los estudios de casos y controles. Consiste en seleccionar para cada caso un control que tenga la misma exposición al posible factor de confusión que queremos controlar. Esta técnica impide medir el efecto de la variable por la que se empareja. El emparejamiento en estudios de casos y controles introduce un sesgo si no se analiza con métodos específicos. Se suele realizar para hacer más eficiente el análisis ajustado en determinadas situaciones, bien mediante estratificación o modelaje multivariable.

c) Restricción

La restricción consiste en seleccionar para el estudio a personas con el mismo nivel de exposición al factor de confusión, para que de esta forma el posible factor de confusión no distorsione la asociación entre la exposición de interés y el efecto. Por ejemplo, si queremos estimar la asociación entre el consumo de café y el cáncer de laringe, podemos seleccionar solo a personas que no fumen o, alternatively, solo a personas que fumen, porque sabemos que el tabaco es un factor de confusión entre el consumo de café y el cáncer de laringe. Al igual que ocurre en el emparejamiento, la restricción conlleva ajustar no solo por el factor de confusión, sino por otra serie de variables asociadas al factor de confusión y posiblemente al desenlace. La desventaja de la restricción es que limita la extrapolación de los resultados a las poblaciones con el mismo nivel de exposición al factor de confusión que la muestra de estudio.

1.5.2. Control de la confusión en la fase de análisis

Durante el análisis podemos controlar el efecto de la confusión mediante estratificación o mediante modelos multivariados. Para ello es necesario que hayamos recogido previamente información de la distribución de las potenciales variables confusoras en nuestra población de estudio.

a) Análisis estratificado

El control de la confusión mediante estratificación consiste en distribuir durante la fase de análisis a los sujetos en subgrupos (estratos) que sean homogéneos en las categorías del factor de confusión. A continuación, calculamos la medida de asociación en cada uno de los estratos, que es independiente del factor de confusión. Por medio de los test de homogeneidad determinamos si las medidas de asociación obtenidas en los estratos son homogéneas o heterogéneas. Si son homogéneas, calculamos una medida de asociación ajustada, esto es, la media ponderada de las medidas de asociación obtenidas en los estratos. Los dos métodos habituales para el cálculo de esta medida de asociación ponderada son la estandarización y la agregación. Los métodos de agregación más utilizados son el método de Woolf y el de Mantel y Haenszel. Existe confusión cuando la medida de asociación ajustada difiere de la medida de asociación cruda (ver capítulo 6).

Veamos un ejemplo hipotético de control de la confusión mediante análisis estratificado.

Se realizó un estudio de cohortes en mujeres con el objetivo de analizar la asociación entre el consumo de grasas en la dieta y la incidencia de cáncer de mama. Tras 20 años de seguimiento se obtuvieron estos resultados (tabla 5).

Tabla 5. Estudio hipotético. Distribución del cáncer de mama en mujeres según el consumo de grasas

	Alto consumo grasas	Bajo consumo grasas
Mujeres con cáncer de mama	72	70
Mujeres-año	60 000	55 000
Tasa de incidencia	0,0012 año ⁻¹	0,0013 año ⁻¹
	En el grupo de mujeres con un ALTO consumo de grasas se diagnosticaron 12 casos de cáncer de mama por cada 10 000 mujeres-año de seguimiento.	En el grupo de mujeres con un BAJO consumo de grasas se diagnosticaron 13 casos de cáncer de mama por cada 10 000 mujeres-año de seguimiento.
Razón de tasas (RT):	0,0012/ 0,0013= 0,92	
	La tasa de incidencia de cáncer de mama es un 8 % menor [(0,92-1)x100] en las mujeres con alto consumo de grasas respecto a las mujeres con bajo consumo de grasas.	

Nos podemos preguntar si el nivel socioeconómico (NSE) puede ser un factor de confusión en la asociación entre el consumo de grasas y el cáncer de mama. Para ello, dividimos la muestra en estratos homogéneos en la exposición al posible factor de confusión y calculamos la medida de asociación en cada uno de los estratos, tal y como se muestra en la tabla 6.

Tabla 6. Estudio hipotético. Distribución del cáncer de mama en mujeres según el consumo de grasas y nivel socioeconómico

Nivel socioeconómico alto				Nivel socioeconómico bajo		
	Alto consumo grasas	Bajo consumo grasas	Total	Alto consumo grasas	Bajo consumo grasas	Total
Cáncer de mama	47 (a_1)	10 (b_1)	57	25 (a_2)	60 (b_2)	85
Mujeres-año	50 000 ($PT1_1$)	15 000 (PTO_1)	65 000 (T_1)	10 000 ($PT1_2$)	40 000 (PTO_2)	50 000 (T_2)
Tasa de incidencia	0,00094	0,00067		0,00250	0,00150	
	En las mujeres con un NSE ALTO y un ALTO consumo de grasas se diagnosticaron 94 casos de cáncer de mama por cada 100 000 mujeres-año de seguimiento.	En las mujeres con un NSE ALTO y un BAJO consumo de grasas se diagnosticaron 67 casos de cáncer de mama por cada 100 000 mujeres-año de seguimiento.		En las mujeres con un NSE BAJO y un ALTO consumo de grasas se diagnosticaron 250 casos de cáncer de mama por cada 100 000 mujeres-año de seguimiento.	En las mujeres con un NSE BAJO y un BAJO consumo de grasas se diagnosticaron 150 casos de cáncer de mama por cada 100 000 mujeres-año de seguimiento.	
Razón de tasas	0,00094/0,00067= 1,41			0,0025/0,0015=1,67		
	Entre las mujeres con NSE ALTO, la tasa de incidencia de cáncer de mama es un 41 % mayor en las que tienen un alto consumo de grasas que en aquellas con bajo consumo de grasas.			Entre las mujeres con NSE BAJO, la tasa de incidencia de cáncer de mama es un 67 % mayor en las que tienen un alto consumo de grasas que en aquellas con bajo consumo de grasas.		

Las medidas de asociación obtenidas en los estratos son independientes del NSE, nuestro posible factor de confusión. En uno de los estratos todas las mujeres tienen un NSE bajo y en el otro estrato todas las mujeres tienen un NSE alto.

Si realizamos un test de homogeneidad de estratos, observamos que la razón de tasas de ambos estratos es homogénea (RT_NSE alto:1,41; RT_NSE bajo:1,67), es decir, que las diferencias entre los estratos son compatibles con variabilidad aleatoria. Por tanto, calculamos una medida de asociación ajustada, es decir, una media ponderada de las medidas de asociación obtenidas en los estratos. En este caso, utilizamos el método de Mantel y Haenszel:

$$RT_{m\_h} = \frac{(\sum a_i * PT_{0i}) / T_i}{(\sum b_i * PT_{1i}) / T_i}$$

$$RT_{m\_h} = \frac{(47 * \frac{15\,000}{65\,000}) + (25 * \frac{40\,000}{50\,000})}{(10 * \frac{50\,000}{65\,000}) + (60 * \frac{10\,000}{50\,000})}$$

$$RT_{m-h} = 1,57$$

Como la RT_{m-h} difiere de la RT_{cruda}, sin ajustar (RT_{m-h}: 1,57; RT_{cruda}: 0,92), concluimos que existe confusión, en este caso de carácter cualitativo, porque ha cambiado la dirección de la medida de asociación. El riesgo de cáncer de mama es un 57% mayor en las mujeres con alto consumo de grasas (media ajustada, una vez eliminada la confusión por el nivel socioeconómico) y no un 8% menor (medida cruda, con presencia de confusión).

Por tanto, se concluye que, la incidencia de cáncer de mama en las mujeres con un alto consumo de grasas es un 57% mayor que la de las mujeres con un consumo bajo, independientemente del nivel socioeconómico.

b) Análisis multivariable

El análisis multivariable es un método analítico que permite, mediante modelos matemáticos, estimar el efecto independiente de varias exposiciones sobre un desenlace, esto es, permite estimar el efecto de una variable de interés controlando por múltiples factores de confusión. Previamente es importante conocer el comportamiento individual de cada uno de los posibles factores de confusión que queremos estimar y no incluir variables que puedan dar lugar a sobreajustes que puedan distorsionar nuestra medida de asociación. (1)

1.6. La confusión residual

A pesar de los esfuerzos que realicemos para hacer un control adecuado de la confusión, nuestras estimaciones pueden estar afectadas, en mayor o menor medida, por la confusión residual, es decir, por aquella confusión que no llegamos a controlar. La confusión residual se puede producir por las siguientes razones: (3)

- Por falta de control de factores de confusión desconocidos.
- Por sesgos de información que hacen que el factor de confusión presente errores de clasificación.
- Porque nuestras variables recojan constructos complejos (ej.: clase social, nivel de estudios, índice de privación económica) y en ocasiones imprecisos en cuanto a su efecto sobre un desenlace.

2. Interacción o modificación del efecto

Decimos que existe interacción cuando la asociación de una exposición con un desenlace es diferente en los distintos estratos de una tercera variable que llamamos modificador de efecto. (6,7)

La interacción no es un sesgo, es un fenómeno natural, que está o no presente independientemente del diseño de nuestro estudio. Por tanto, a diferencia de la confusión, que identificamos para tratar de controlarla, la interacción la identificamos para describir lo que ocurre en cada uno de los estratos de la variable modificadora de efecto.

Veamos un ejemplo hipotético:

Mediante un estudio de casos y controles queremos evaluar la relación potencial entre el consumo de alcohol y el cáncer de vejiga. Los resultados que obtenemos se muestran en la tabla 7.

Tabla 7. Estudio hipotético. Distribución del cáncer de vejiga según el consumo de alcohol

		Consumo de alcohol	
		SÍ	NO
Cáncer de vejiga	SÍ	233	128
	NO	297	360
Odds Ratio (OR):		2,21. La probabilidad de desarrollar cáncer de vejiga en personas que consumen alcohol es 2,21 veces la probabilidad correspondiente de las que no consumen alcohol.	

Realizamos un análisis estratificado (tabla 8) para analizar el posible efecto de la región de origen en la asociación consumo de alcohol-cáncer de vejiga y obtenemos estos resultados:

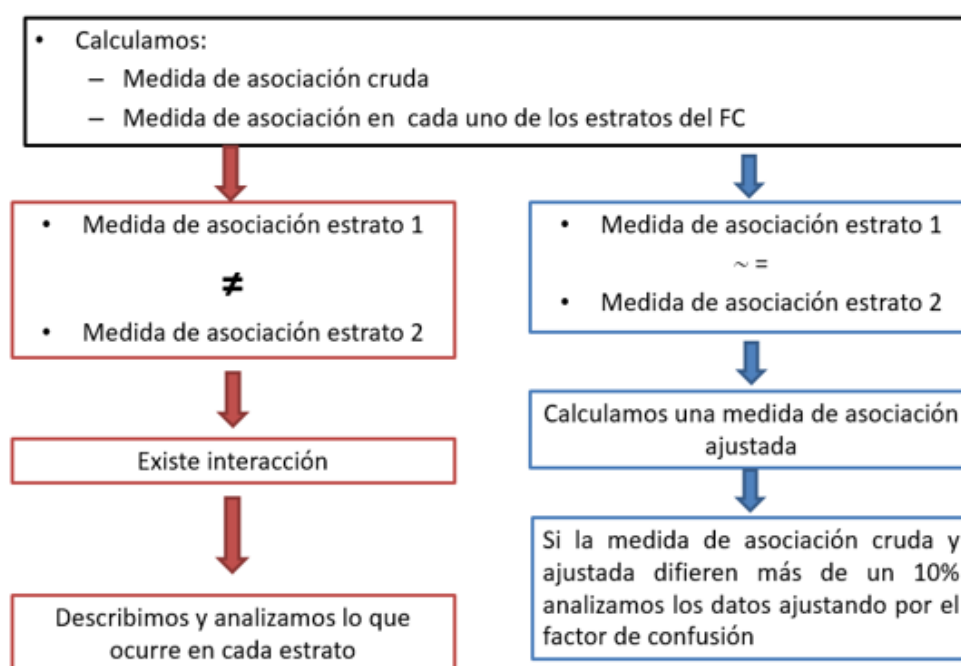
Tabla 8. Estudio hipotético. Distribución del cáncer de vejiga según el consumo de alcohol y región de origen

Población europea		Consumo de alcohol	
		SÍ	NO
Cáncer de vejiga	SÍ	72	41
	NO	106	105
Odds Ratio:		1,74. En la población europea la probabilidad de desarrollar cáncer de vejiga en personas que consumen alcohol es 1,74 veces la probabilidad correspondiente de las que no consumen alcohol.	
Población africana		Consumo de alcohol	
		SÍ	NO
Cáncer de vejiga	SÍ	93	54
	NO	113	113
Odds Ratio:		1,72. En la población africana la probabilidad de desarrollar cáncer de vejiga en personas que consumen alcohol es 1,72 veces la probabilidad correspondiente de las que no consumen alcohol.	
Población asiática		Consumo de alcohol	
		SÍ	NO
Cáncer de vejiga	SÍ	68	33
	NO	78	142
Odds Ratio:		3,75. En la población asiática la probabilidad de desarrollar cáncer de vejiga en personas que consumen alcohol es 3,75 veces la probabilidad correspondiente de las que no consumen alcohol.	

Si observamos los estratos y realizamos un test de homogeneidad, vemos que las medidas de asociación de los estratos son heterogéneas y que la magnitud de la asociación del consumo de alcohol con el cáncer de vejiga es mayor en la población asiática que en las poblaciones africana o europea. En este caso decimos que existe interacción y que el lugar de origen es un modificador de efecto de la asociación previamente descrita. Una vez identificada una interacción, describimos la asociación estimada en cada uno de los estratos.

Figura 1: Resumen modificador de efecto vs. factor de confusión

INTERACCIÓN	CONFUSIÓN
Existe interacción cuando el efecto de una exposición sobre la enfermedad es diferente para los distintos estratos de una tercera variable: $OR_{\text{estrato 1}} \neq OR_{\text{estrato 2}} (OR/RR/RT..)$	La confusión es una distorsión de la asociación estimada entre una exposición y el efecto debido a la presencia de una tercera variable. Esta tercera variable la llamamos factor de confusión. $OR_{\text{crudo}} \neq OR_{\text{ajustado}} (OR/RR/RT..)$
La interacción no es un sesgo, es un fenómeno natural, que está presente o no independientemente del diseño de nuestro estudio.	La confusión es un sesgo, que puede estar presente independientemente del diseño de nuestro estudio, excepto, si acaso, en los estudios de intervención aleatorizados con tamaño muestral suficiente.
Cuando existe interacción, no debemos obtener un estimador global del efecto: $OR_{\text{estrato 1}} \neq OR_{\text{estrato 2}} (OR/RR/RT..)$	En presencia de confusión, debemos obtener un estimador global del efecto, que es la medida de asociación ajustada: $OR_{\text{estrato 1}} \approx OR_{\text{estrato 2}} (OR/RR/RT..)$ → OR ajustado
Las diferencias en las medidas de asociación entre los estratos podemos identificarlas mediante un test estadístico de homogeneidad.	La presencia de la confusión depende de las diferencias en la magnitud de las medidas de asociación cruda y ajustada, independientemente de la probabilidad de que tales diferencias se deban o no al azar.
Cuando existe interacción, los resultados son los estimadores de la asociación en los estratos conformados por el factor modificador del efecto.	Cuando existe confusión, los resultados son el estimador global agregado, mediante la media ponderada de los estimadores en los estratos del factor de confusión.

Figura 2: Identificación de la interacción y la confusión mediante análisis estratificado.

3. Referencias

1. de Irala J, Martínez-González MÁ, Grima FG. ¿Qué es una variable de confusión? Med Clin (Barc). 2001;117:377-85.
2. de Irala J, Martínez-González MA. Errores en la estimación de medidas de asociación en estudios epidemiológicos. En: de Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. Barcelona: Ariel Ciencias Médicas, 2008:257-346.
3. Villar Álvarez F. Sesgos y factores de confusión. En: Royo Bordonada MÁ, Damián Moreno J. Manual del Método Epidemiológico. Madrid: Escuela Nacional de Sanidad, 2009:47-74.
4. Szklo M & Nieto FJ. Epidemiology: beyond the basics. Fourth edition. Burlington: Jones & Bartlett Learning, 2019.
5. Delgado Rodríguez M, Llorca Díaz J, Martínez-González MA. Sesgos. En: Piédrola G. Medicina Preventiva y Salud Pública. Barcelona: Elsevier, 2015.
6. de Irala J, Martínez-González MA. Variables Modificadoras del Efecto. En: de Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. Barcelona: Ariel Ciencias Médicas, 2008:346-70.

7. de Irala J, Martínez-González M Á, Grima FG. ¿Qué es una variable modificadora del efecto? Med Clin (Barc). 2001;117:297-302.

Capítulo 6.

Análisis de datos epidemiológicos

JAVIER DAMIÁN y NURIA ARAGONÉS SANZ

1. Fundamentos del análisis

1.1. Pruebas de significación estadística

2. Análisis crudo

2.1. Datos con denominadores de personas-tiempo

2.2. Datos de recuento

3. Análisis estratificado

3.1. Estandarización

3.2. Métodos de agregación. Métodos de Woolf y Mantel-Haenszel

3.3. Diferencias entre los métodos de análisis estratificado

3.4. Confusión residual

3.5. Aplicación al análisis de datos emparejados

4. Referencias

El análisis de datos consiste en la obtención de medidas resumen que ayuden a realizar inferencias. En la mayoría de las situaciones, el interés reside en la estimación de parámetros. En ocasiones el interés se centra en el contraste de hipótesis nulas.

Cuando el objetivo del análisis es obtener información cuantitativa sobre el valor de algún parámetro poblacional, como una media, una prevalencia o una razón de tasas, se habla de realizar una *estimación*, que a su vez puede ser de dos tipos, puntual y por intervalo de confianza. Mediante la *estimación puntual* se obtiene un único valor del parámetro desconocido (por ejemplo, una razón de tasas de 2,43). Asumiendo que no hay sesgos, la estimación puntual es la mejor estimación numérica de un efecto a partir de nuestros datos. Sin embargo, las estimaciones obtenidas en procesos de muestreo están afectadas por error aleatorio, por lo que normalmente es necesario ofrecer un rango de valores del parámetro. La *estimación por intervalo de confianza* persigue obtener un rango de las hipótesis más compatibles con el estimado puntual. Este rango será más amplio cuanto mayor sea el error aleatorio. Se dice que una estimación tiene gran *precisión* cuando el intervalo de confianza es estrecho.

Como se ha mencionado, el proceso de estimación de parámetros puede verse afectado por una serie de errores, de diversa naturaleza y magnitud, dependiendo del diseño del estudio. Estos errores pueden ser sistemáticos o aleatorios. Los errores sistemáticos incluyen los distintos tipos de sesgos y factores de confusión, y es posible predecir su dirección, aunque es complicado calcular su magnitud. Los errores aleatorios dependen fundamentalmente de la variabilidad del fenómeno que se mide y del tamaño muestral (y de determinados errores de medida). En los errores aleatorios se puede medir su magnitud pero no su dirección.

Para estimar la magnitud del error aleatorio asociado a las mediciones se utilizan métodos estadísticos. En este capítulo se mostrarán métodos convencionales, *frecuentistas o de muestreos reiterados*. Estos se basan en el comportamiento que se supone tendrían las mediciones si el estudio lo realizáramos innumerables veces en las mismas condiciones y las diferencias que encontráramos en sus resultados fueran debidas únicamente a variabilidad aleatoria. Se basan en unas asunciones difíciles de cumplir en la mayoría de los estudios observacionales y en muchos experimentales, (1) pero aquí se describen debido a que su utilización es muy generalizada y, sobre todo, porque nos dan una idea aproximada de la magnitud del error aleatorio, cuya medición es uno de los objetivos de este capítulo.

En este capítulo se presentan algunos procedimientos estadísticos aplicados al análisis de datos de estudios epidemiológicos cuyo objetivo es evaluar la asociación, causal o no, entre un determinante (característica, exposición, hábito...) y la ocurrencia de una enfermedad en una población diana. En primer lugar se aborda

la estimación de efectos para el caso del análisis crudo o simple. En segundo lugar se considera el análisis estratificado, en el que se tienen en cuenta otras variables además de las de exposición y desenlace.

Los cálculos que se deben realizar en el análisis de datos dependen del modelo estadístico que se utilice. En general, y debido a su facilidad, se utilizarán modelos basados en la distribución normal. Estos métodos, también denominados *aproximados* o asintóticos, son apropiados en el caso de muestras suficientemente grandes. Con muestras pequeñas, los cálculos deben realizarse con métodos exactos.

1. Fundamentos del análisis

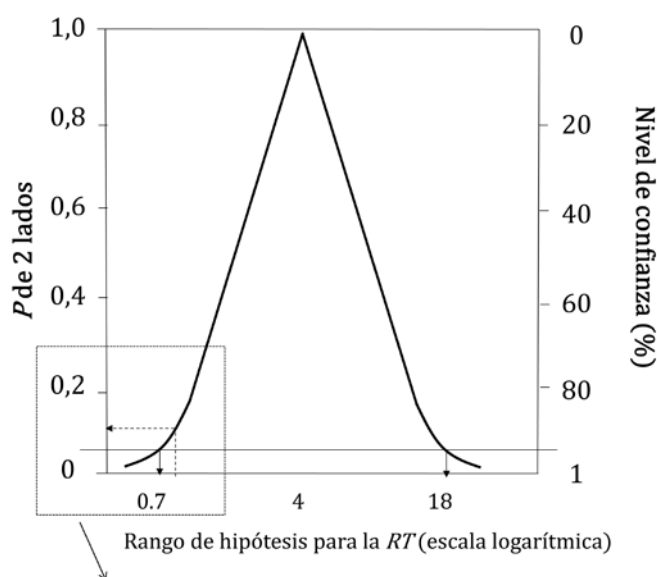
Los resultados de un análisis epidemiológico se deben situar en un marco de trabajo en el cual, normalmente, se selecciona una muestra con el objeto de estimar un parámetro poblacional. Esto implica que, por la naturaleza propia del muestreo, el verdadero valor del parámetro siempre será desconocido. Solo se puede conjeturar (inferir) sobre su valor, que por tanto es hipotético. Intuitivamente se entiende que hay hipótesis sobre el verdadero valor más compatibles con el resultado empírico, “muestral”, aquellas de valor similar al resultado empírico; y otras hipótesis menos compatibles, aquellas que se alejan mucho del valor estimado. Lo que hacemos es una valoración cuantitativa de la compatibilidad de nuestro resultado muestral con distintas hipótesis. Una forma de hacerlo es mediante el valor P asociado a cada hipótesis, que es un valor entre 0 y 1 calculado bajo un modelo estadístico sujeto a ciertas asunciones, entre ellas que la hipótesis es cierta. A mayor valor P más compatible es nuestro resultado con la hipótesis. Es el método habitual usado en la inferencia estadística frecuentista. La distribución de todos estos valores P se denomina *función de valores P* (2,3) (Figura 1). Conviene insistir en que cada uno de los valores P se calcula asumiendo que la hipótesis contrastada fuera verdad, por lo que no debe entenderse un valor P como la probabilidad de que esa hipótesis sea verdad. En esta función de valores P están contenidos los elementos básicos del análisis que se utilizan habitualmente en inferencia estadística frecuentista, como por ejemplo los intervalos de confianza.

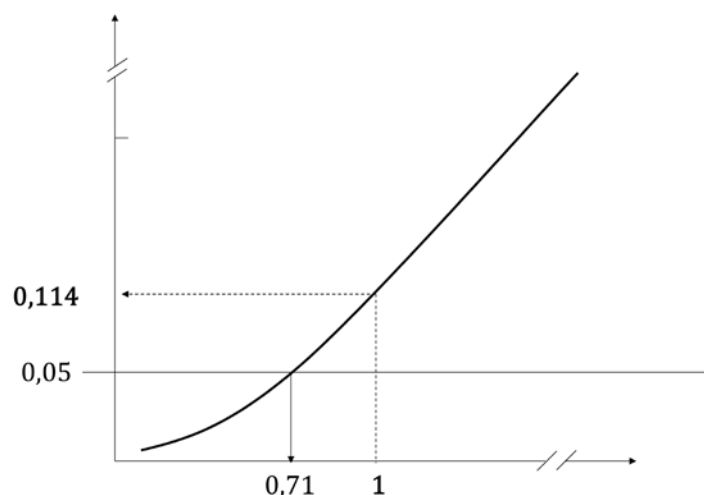
Hay dos tipos de valores P : unilaterales y bilaterales. Entre los unilaterales hay dos tipos: menor (P -menor) y mayor (P -mayor). El valor P unilateral asociado a una hipótesis se calcula como la probabilidad de obtener —en repeticiones ilimitadas del estudio en las mismas condiciones— un resultado igual o mayor (o igual o menor si es una P -menor) que el observado en nuestro estudio si la hipótesis que

contrastamos fuera verdad y no hubiera sesgo. El valor P bilateral es el doble del valor de la P unilateral más baja.

La función de valores P , además de ofrecer una idea general del error aleatorio de la estimación, permitirá calcular algunos valores de interés, como los límites de los diferentes intervalos de confianza. Un intervalo de confianza al $100(1-\alpha)\%$ incluye todos los valores para los cuales su correspondiente valor P bilateral excede un determinado valor α elegido (valores entre 0 y 1, habitualmente 0,05, 0,1 o 0,01), es decir, es un rango que incluye los valores asociados a las hipótesis más compatibles con el estimado puntual medido en nuestro estudio. El valor P más alto de la función es el correspondiente a la hipótesis que coincide con nuestra estimación puntual. Conforme las hipótesis a contrastar se van alejando, sus valores P van siendo menores. Las dos hipótesis cuyo valor P bilateral es 0,1 marcan los límites inferior y superior de un intervalo de confianza del 90 %, y las dos hipótesis cuyo valor P bilateral es 0,05 marcan los límites de un intervalo de confianza del 95 %. En una hipotética repetición ilimitada del estudio en las mismas condiciones, los intervalos de confianza contendrán el verdadero parámetro con una frecuencia no menor que el nivel de confianza —si el modelo estadístico es correcto y no hay sesgos—. Se asume, por tanto, que la única razón de una diferencia entre nuestro valor observado y el verdadero valor del parámetro es la variabilidad aleatoria. Esta interpretación depende de asunciones de muestreo y modelos de estadísticos que no se cumplen en la mayoría de los estudios epidemiológicos, por lo que conviene tomar los intervalos de confianza como una indicación aproximada del error aleatorio asociado a la medición y no como un elemento clave en toma de decisiones.

Figura 1. Función de valores P correspondiente a los datos del ejemplo 1. El pico de la curva corresponde a la hipótesis que representa el estimado puntual ($RT=4,23$). Zona ampliada: El valor P (exacto) para la hipótesis nula ($RT=1$) es 0,114. Los límites (exactos) del intervalo de confianza al 95 % son 0,71 y 18,5.





1.1. Pruebas de significación estadística

En muchos escenarios se puede dedicar especial atención a una hipótesis que representa la ausencia de asociación (en sentido amplio). Es la hipótesis nula, una de las infinitas hipótesis contrastables en la función de valores P . Este particular valor P es una medida de la compatibilidad de los datos obtenidos en el estudio —en el ejemplo de la figura 1 es una RT de 4,23— con esta hipótesis ($RT=1$). En este ejemplo el valor P bilateral asociado a la hipótesis nula es de 0,114.

Las pruebas de significación estadística consisten en determinar si valores P asociados a hipótesis nulas son menores o mayores que un nivel preestablecido, en cuyo caso se descarta o no la hipótesis nula. Queda claro que realizar este tipo de pruebas de significación estadística solo se entiende en un contexto de decisión en el que se ha de optar por un curso de acción u otro en función del resultado. Por tanto debe explicitarse dicho esquema de decisión en el que el nivel elegido sea coherente con la gravedad de una eventual equivocación y, en caso de optar por una hipótesis alternativa, ésta debe plantearse antes del estudio. En todo caso nunca debe basarse una decisión únicamente en el resultado de estas pruebas, de dudosa utilidad.

2. Análisis crudo

Se entiende como análisis crudo aquel que tiene en cuenta solo las variables principales, normalmente exposición y efecto. Estos análisis se usan cuando no es necesario tener en cuenta ningún factor más aparte de la exposición y de la enfer-

medad de interés, como ocurre a menudo en estudios experimentales, en los que la asignación aleatoria de la exposición suele prevenir la confusión. Además, el análisis crudo debe ser siempre el primer paso antes de realizar análisis que incorporen otras variables, como ocurre en los estudios no experimentales, que normalmente requieren otros métodos, como el análisis estratificado o el modelaje multivariante, para poder tener en cuenta el papel de otros factores además de las variables principales (factores de confusión y variables de interacción).

Para facilitar su descripción, los métodos de análisis que se presentan a continuación se limitan al caso de una variable de exposición y una de resultado, ambas dicotómicas.

2.1. Datos con denominadores de personas-tiempo

En un estudio de seguimiento con personas-tiempo como unidades básicas de observación, la información utilizada en el análisis normalmente se presenta en la siguiente tabla de notación:

Tabla 1. Notación de un estudio de seguimiento con una variable de exposición dicotómica y denominadores persona-tiempo

	Expuestos	No expuestos
Casos	a	b
Personas-tiempo	PT_1	PT_0

- *Diferencia de tasas* de incidencia:

$$DT = I_1 - I_0 = \frac{a}{PT_1} - \frac{b}{PT_0}$$

La fórmula para obtener el intervalo de confianza:

$$DT \pm Z_{\alpha/2} EE(DT)$$

$Z_{\alpha/2}$ es el valor de la distribución normal estándar correspondiente al nivel de confianza elegido. Por ejemplo, para un 90 % es 1,65; para un 95 % es 1,96; y para un 99 % es 2,58. Y $EE(DT)$ es el error estándar de la diferencia de tasas:

$$EE(DT) = \sqrt{\frac{a}{PT_1^2} + \frac{b}{PT_0^2}}$$

- *Razón de tasas:*

$$RT = \frac{I_1}{I_0} = \frac{a}{PT_1} / \frac{b}{PT_0}$$

Dado que la razón de tasas de incidencia no tiene una distribución simétrica (Figura 2), es conveniente emplear una transformación logarítmica, que permite usar la distribución normal. De esta forma, se calculan límites de confianza simétricos para el logaritmo de la razón de tasas, siendo necesario posteriormente deshacer la transformación mediante la función inversa, la exponencial:

$$\exp[\ln RT \pm Z_{\alpha/2} EE(\ln RT)]$$

$$EE(\ln RT) = \sqrt{\frac{1}{a} + \frac{1}{b}}$$

Ejemplo 1. En el estudio de seguimiento del NHANES II (National Health and Nutrition Examination Survey), se han obtenido los siguientes datos de mortalidad cardiovascular en el grupo de edad de 30 a 49 años:

Tabla 2. Mortalidad cardiovascular en el grupo de edad de 30 a 49 años

	Hipertensos	Normotensos
Casos	3	7
Personas-año	2868	28 309

$$DT = \frac{3}{2868 \text{ p-a}} - \frac{7}{28\,309 \text{ p-a}} = 0,00080 \text{ año}^{-1}$$

$$EE(DT) = \sqrt{\frac{3}{2868^2} + \frac{7}{28\,309^2}} = 0,00061 \text{ año}^{-1}$$

Y el intervalo de confianza al 90 % expresado por 10^5 personas-año:

$$80 \pm 1,65 \times 61 = -21 \text{ a } 181 \text{ por } 10^5 \text{ personas-año}$$

De forma equivalente para la RT :

$$RT = \frac{3/2868}{7/28\,309} = 4,23$$

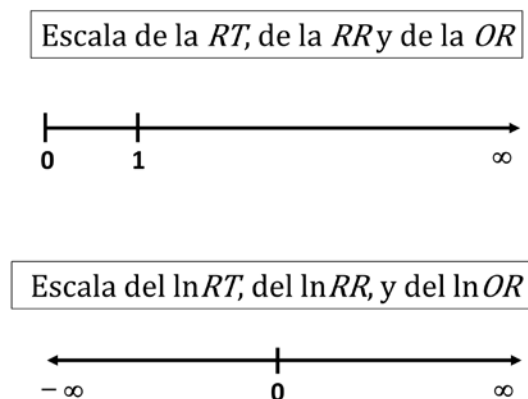
$$EE(\ln RT) = \sqrt{\frac{1}{3} + \frac{1}{7}} = 0,69$$

Y el intervalo de confianza al 90 % de la RT :

$$\exp[\ln(4,23) \pm 1,65 \times 0,69] = 1,35 \text{ a } 13,21$$

(Si el intervalo de la RT es al 95%: $\exp[\ln(4,23) \pm 1,96 \times 0,69] = 1,10 \text{ a } 16,4$)

Figura 2. Rango de valores de la razón de tasas, razón de riesgos y odds ratio, en escala original y en escala logarítmica



2.2. Datos de recuento

La información básica, tanto para estudios de cohortes con denominadores en forma de personas en riesgo como para estudios de prevalencia y de casos y controles, puede organizarse en forma de una tabla de 2×2 , en la que las 4 celdas son frecuencias de sujetos clasificados de acuerdo con la presencia o ausencia de exposición y de enfermedad.

Tabla 3. Notación de estudios con datos de recuento, y variables de exposición y enfermedad dicotómicas

	Expuestos	No expuestos
Casos	a	b
No casos	c	d
Total	N_1	N_0

- *Diferencia de riesgos:*

$$DR = IA_1 - IA_0 = \frac{a}{N_1} - \frac{b}{N_0}$$

Intervalo de confianza:

$$DR \pm Z_{\alpha/2} EE(DR)$$

$$EE(DR) = \sqrt{\frac{ac}{N_1^3} + \frac{bd}{N_0^3}}$$

- *Razón de riesgos:*

$$RR = \frac{IA_1}{IA_0} = \frac{a}{N_1} / \frac{b}{N_0}$$

Como en el caso de las razones de tasas, para el cálculo de los intervalos de confianza de la razón de riesgos también se emplea la transformación logarítmica.

$$\exp[\ln RR \pm Z_{\alpha/2} EE(\ln RR)]$$

$$EE(\ln RR) = \sqrt{\frac{c}{aN_1} + \frac{d}{bN_0}}$$

- ♦ Para estudios en los que se calculan odds de exposición (casos y controles) o de enfermedad (cohortes), la medida de interés es la *odds ratio*:

$$OR = \frac{ad}{bc}$$

En este caso, el cálculo del error estándar y del intervalo de confianza se realiza utilizando también la transformación logarítmica:

$$\exp[\ln OR \pm Z_{\alpha/2} EE(\ln OR)]$$

$$EE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Ejemplo 2. (datos del ejemplo 11 del capítulo 2). Estudio de intervención para medir la efectividad de una intervención nutricional personalizada en pacientes hospitalizados en 6 centros. El grupo control recibió la alimentación habitual del hospital. El seguimiento fue de 30 días. Los resultados en mortalidad por todas causas fueron:

Tabla 4. Mortalidad a 30 días por todas las causas según intervención

	Intervención	Control
Casos	73	100
No casos	942	913
Total	1015	1013

Los resultados son los siguientes:

$$IA_{\text{intervención}} = 0,072; IA_{\text{control}} = 0,099$$

$$DR = -0,027; EE(DR) = 0,0124; \quad IC90\%: -0,006 \text{ a } -0,047$$

$$RR = 0,73; EE(\ln RR) = 0,1474; \quad IC90\%: 0,57 \text{ a } 0,93$$

$$OR = 0,71; EE(\ln OR) = 0,1608; \quad IC90\%: 0,54 \text{ a } 0,92$$

3. Análisis estratificado

El análisis estratificado engloba un conjunto de métodos que se basan en la creación de subgrupos con el objeto de abordar determinados problemas, como la identificación y control de factores de confusión y la evaluación de la modificación del efecto.

La *confusión*, como se ha explicado en el capítulo anterior, es una distorsión en la medida de asociación entre una exposición y una enfermedad resultante del efecto de variables determinantes de la enfermedad asociadas de forma independiente con la exposición.

La *modificación del efecto* (o *interacción estadística*), también definida en el capítulo previo, se refiere al cambio en la magnitud de la asociación entre la exposición y la enfermedad en función del valor de otras variables llamadas modificadoras de efecto.

La estrategia general para examinar la existencia de confusión o de modificación de efecto consiste en crear subgrupos (también llamados estratos), formados por las categorías de la o las variables confusoras (o modificadoras de efecto), y medir el efecto (razón de riesgos, odds ratio, diferencia de tasas...) en dichos subgrupos. Por ejemplo, calcular la *RR* en grupos de edad, en hombres y en mujeres, o en grupos de nivel socioeconómico (por ejemplo, bajo, medio y alto). Al ser grupos homogéneos de individuos en cuanto a las variables de confusión que los definen, el efecto medido en cada subgrupo no podrá ser atribuido a esos factores de confusión por lo que serán estimaciones no confundidas o ajustadas por esos factores. Si no se observan diferencias relevantes (más allá de lo esperable por va-

riabilidad aleatoria) en el efecto entre los subgrupos, se puede proceder a calcular la media de los valores estimados en los subgrupos, para obtener un valor único no confundido por las variables de estratificación. Cuando el valor medio del efecto en los subgrupos es diferente del valor crudo, es posible que se exista confusión por alguna o algunas de las variables de estratificación. Si se observan diferencias en los valores estimados en los subgrupos, más allá de lo que se considere como variabilidad aleatoria, puede que estemos ante un fenómeno de modificación del efecto.

Una de las diferencias más importantes entre la confusión y la modificación del efecto es que la confusión es un sesgo y, por tanto, hay que identificarlo y controlarlo, por lo que procederá calcular el promedio del efecto no confundido observado en los subgrupos. Sin embargo la modificación del efecto no es un sesgo, es un atributo del fenómeno que se estudia, por lo que el objetivo es describirlo y evaluarlo. Si se aprecia modificación del efecto lo esencial es describir e interpretar el patrón observado en los subgrupos, y en general no procede calcular un estimador promedio.

La obtención de estimados ajustados se basa en el cálculo del *promedio ponderado* de los valores estimados de los subgrupos. Existen varios métodos en función de la elección de las ponderaciones (pesos) que se asignarán a cada subgrupo. Aquí se presentan los dos métodos más habituales: la estandarización y la agregación (en inglés “pooling”). Y entre los métodos de agregación el *método de Woolf* y el método de *Mantel y Haenszel*.

Los métodos de agregación asumen homogeneidad de los efectos en los subgrupos. Es decir, asumen que en cada subgrupo se está estimando el mismo efecto y que las diferencias encontradas son debidas a variabilidad aleatoria. El *método de Woolf* utiliza pesos proporcionales al inverso de la varianza de cada subgrupo y el *método de Mantel-Haenszel* utiliza un grupo especial de pesos basados en los datos de cada subgrupo, con un componente igualmente proporcional al error aleatorio de cada subgrupo. La *estandarización* utiliza unos pesos que pueden ser incluso externos a los datos y no asume homogeneidad de efectos de los subgrupos. A continuación se describen estos métodos.

3.1. Estandarización

La estandarización es un procedimiento de control de factores de confusión que se basa en la sustitución de las distribuciones de los factores de confusión de las poblaciones que se comparan por otras distribuciones estándar (o comunes). Se muestra como ejemplo la estandarización de tasas, pero el procedimiento es equivalente en el caso de incidencias acumuladas, prevalencias y otros parámetros.

En el ejemplo se muestra la estandarización por una sola variable (la edad) pero el procedimiento puede incluir más de una.

Tabla 5. Notación de un estudio de seguimiento con tasas de incidencia por subgrupos de edad

Grupo de edad	N.º de casos (a_i)	Personas-tiempo (PT_i)	Tasa de incidencia específica ($I_i = a_i / PT_i$)
30-49	a_1	PT_1	I_1
50-69	a_2	PT_2	I_2
70+	a_3	PT_3	I_3
Total	a	PT	I

La tasa de incidencia total (también tasa cruda o bruta) es una *media ponderada* de las tasas específicas por subgrupos de la población exhaustivos y mutuamente excluyentes, donde el peso de cada subgrupo es su tamaño. En el caso de una sola variable, por ejemplo la edad:

$$I = \frac{a}{PT} = \frac{a_1 + a_2 + a_3}{PT_1 + PT_2 + PT_3} = \frac{I_1 PT_1 + I_2 PT_2 + I_3 PT_3}{PT_1 + PT_2 + PT_3} = \frac{\sum I_i PT_i}{\sum PT_i}$$

Si los pesos se expresan como proporción de la suma total de pesos, el denominador es 1 y la media ponderada se reduce a la suma de las tasas específicas multiplicadas por su peso en forma de proporción:

$$I = \left(I_1 \frac{PT_1}{PT} \right) + \left(I_2 \frac{PT_2}{PT} \right) + \left(I_3 \frac{PT_3}{PT} \right)$$

Se puede apreciar que la tasa cruda, I , depende no solo de la magnitud de las tasas específicas sino también de los pesos de cada grupo.

Cálculo de la tasa estandarizada. Si se sustituyen los pesos PT_i (la distribución por edad) de esta población por los de otra población (w_i) —normalmente llamada “estándar” o de referencia— se obtiene una *tasa estandarizada*:

$$I_E = \frac{\sum I_i w_i}{\sum w_i}$$

Igualmente se pueden utilizar los pesos de la distribución estándar en forma de proporción:

$$I_E = \sum I_i \frac{w_i}{\sum w_i}$$

Por tanto, una tasa estandarizada es una medida hipotética. Corresponde a la tasa que tendría la población de estudio si en lugar de tener su distribución por edad tuviera la distribución por edad estándar. Esta tasa hipotética no tiene utilidad si no es para ser comparada con otras tasas que hayan sido estandarizadas utilizando la misma distribución por edad. Si calculamos la tasa estandarizada para *otra población* utilizando el *mismo grupo de pesos* (la misma distribución estándar), las diferencias entre las dos tasas estandarizadas no podrán ser atribuidas a una diferente distribución por edad. Así, las tasas estandarizadas de dos poblaciones, una de expuestos (I_1) y otra de no expuestos (I_0), utilizando la misma distribución por edad (w_i), serán:

$$I_{1E} = \frac{\sum I_{1i} w_i}{\sum w_i}; \quad I_{0E} = \frac{\sum I_{0i} w_i}{\sum w_i}$$

Nótese que ambas tasas han utilizado el mismo grupo de pesos (w_i). Estos pesos pueden venir de diversas fuentes: de la distribución de una de las dos poblaciones que se comparan, de la suma de las dos poblaciones, o de otra distribución (real o no). En principio, en lo que se refiere a efectos del ajuste, lo importante es que la distribución sea la misma para ambas poblaciones, aunque es recomendable utilizar distribuciones estándar asimilables a alguna población reconocible sobre la que se quiera aplicar los resultados.

Para obtener medidas de efecto estandarizadas, se calcula la diferencia o la razón entre las dos tasas estandarizadas:

$$DT_E = I_{1E} - I_{0E} = \frac{\sum I_{1i} w_i}{\sum w_i} - \frac{\sum I_{0i} w_i}{\sum w_i}$$

$$RT_E = \frac{I_{1E}}{I_{0E}} = \frac{\sum I_{1i} w_i}{\sum w_i} \bigg/ \frac{\sum I_{0i} w_i}{\sum w_i}$$

La DT estandarizada puede expresarse como un promedio ponderado de las diferencias de tasas específicas con los mismos pesos que se usen para la estandarización de las tasas:

$$DT_E = \frac{\sum DT_i w_i}{\sum w_i}$$

Sin embargo, esto no ocurre con la razón de tasas (ni de riesgos). La RT estandarizada se puede expresar como un promedio ponderado de las razones de tasas específicas, RT_i , donde los pesos son el producto de las tasas específicas de los no expuestos multiplicadas por los pesos de la estandarización:

$$RT_E = \frac{\sum RT_i I_{0i} w_i}{\sum I_{0i} w_i}$$

En estos ejemplos se ha utilizado la edad como variable de estandarización, pero el mismo proceso puede utilizarse para estandarizar por cualquier otra variable o combinación de variables: por sexo, por clase social; por edad y sexo; por edad, sexo y clase social, etc. En el ejemplo se comparan solo dos poblaciones pero la estandarización se puede utilizar para comparar múltiples poblaciones.

Los procedimientos son similares para el cálculo de incidencias acumuladas o prevalencias estandarizadas. Se sustituyen en las fórmulas anteriores las tasas de incidencia I por las incidencias acumuladas IA_i o prevalencias P_i , y las PT_i por los denominadores de las incidencias acumuladas o prevalencias, N_i .

Ejemplo 3. Mortalidad por enfermedad cardiovascular en el estudio de seguimiento NHANES II. 7801 personas libres de enfermedad cardiovascular basal seguidas un promedio de 15 años.

Tabla 6. Mortalidad cardiovascular en el estudio NHANES II

Edad	Hipertensos		Normotensos	
	Casos	Personas-año	Casos	Personas-año
30-49	3	2868	7	28 309
50-69	80	13 867	67	31 016
70 +	253	14 842	172	15 348
Total	336	31 577	246	74 673

Las tasas crudas son:

$$I_1 = 336/31\,577 \text{ p-a} = 0,01064 \text{ año}^{-1} \quad I_0 = 246/74\,673 \text{ p-a} = 0,00329 \text{ año}^{-1}$$

Y la razón de tasas crudas: $RT = I_1/I_0 = 3,2$.

Se puede apreciar que el grupo de hipertensos tiene una distribución claramente más envejecida (9 %, 44 %, y 47 %, para los más jóvenes, medianos y mayores respectivamente) comparada con la de normotensos (38 %, 42 %, y 21 %). Al estar la mortalidad por enfermedad cardiovascular fuertemente asociada con la edad, una parte del efecto encontrado en la comparación de tasas crudas ($RT=3,2$) podría ser atribuible a la diferencia de edad.

Las tasas específicas por edad no están afectadas por este problema ya que en cada grupo se asume que no hay diferencia de edad (en este ejemplo, por simplicidad se plantean solo 3 grupos y podría quedar alguna diferencia lo que dejaría cierta confusión residual). Las tasas específicas por edad junto con las razones de tasas específicas son:

$$I_{1\,30-49} = 3/2868 = 0,00105 \text{ año}^{-1}; I_{0\,30-49} = 7/28\,309 = 0,00025 \text{ año}^{-1}; (RT_{30-49} = 4,23)$$

$$I_{1\,50-69} = 80/13\,867 = 0,00577 \text{ año}^{-1}; I_{0\,50-69} = 67/31\,016 = 0,00216 \text{ año}^{-1}; (RT_{50-69} = 2,67)$$

$$I_{1\,70+} = 253/14\,842 = 0,01705 \text{ año}^{-1}; I_{0\,70+} = 172/15\,348 = 0,01121 \text{ año}^{-1}; (RT_{70+} = 1,52)$$

Tomando como grupo de pesos estándar la suma de ambas poblaciones:

$$w_{30-49} = 2868 + 28\,309 = 31\,177 \text{ p-a.}$$

$$w_{50-69} = 13\,867 + 31\,016 = 44\,883 \text{ p-a.}$$

$$w_{70+} = 14\,842 + 15\,348 = 30\,190 \text{ p-a.}$$

Las tasas estandarizadas son:

$$I_{1E} = \frac{\sum I_{1i}w_i}{\sum w_i} = \frac{(0,00105 \times 31\,177) + (0,00577 \times 44\,883) + (0,01705 \times 30\,190)}{31\,177 + 44\,883 + 30\,190} = 0,00759 \text{ año}^{-1}$$

$$I_{0E} = \frac{\sum I_{0i}w_i}{\sum w_i} = \frac{(0,00025 \times 31\,177) + (0,00216 \times 44\,883) + (0,01121 \times 30\,190)}{31\,177 + 44\,883 + 30\,190} = 0,00417 \text{ año}^{-1}$$

Y la razón de tasas estandarizadas:

$$RT_E = \frac{I_{1E}}{I_{0E}} = \frac{0,00759 \text{ año}^{-1}}{0,00417 \text{ año}^{-1}} = 1,82$$

Que se interpreta como la razón de tasas esperable si ambas poblaciones tuvieran la misma distribución por edad estándar. Es decir, es una razón de tasas ajustada por edad. En todo caso, hay que recordar la importancia de describir cualquier patrón observado en los subgrupos, como en este ejemplo donde se observa que la asociación es más clara en los más jóvenes y muestra una tendencia decreciente con la edad. Por otra parte, si se calcula el impacto absoluto (mediante la diferencia de tasas), éste es mucho más elevado en el grupo de mayor edad.

3.1.1. Razón de mortalidad estandarizada

Es muy habitual comparar una población índice o expuesta con una población de referencia y utilizar la distribución de la población índice como estándar. En estas situaciones la tasa estandarizada de la población expuesta coincide con su tasa cruda, por lo que la razón de tasas estandarizadas se convierte en la razón entre un número de *casos observados* en la población expuesta y un número de *casos esperados*: los casos que sería esperable obtener si la población expuesta tuviera las tasas de la población de referencia. Esta razón ha sido denominada *Razón de Mortalidad (o morbilidad) Estandarizada (RME)*. Los pesos estándar son los de la población expuesta,

$$w_i = PT_{1i}$$

$$RME = \frac{\sum I_{1i}PT_{1i}}{\sum PT_{1i}} \bigg/ \frac{\sum I_{Ri}PT_{1i}}{\sum PT_{1i}} = \frac{\sum I_{1i}PT_{1i}}{\sum I_{Ri}PT_{1i}} = \frac{\text{total casos observados}}{\text{total casos esperados}}$$

I_{Ri} es el conjunto de tasas específicas de la población de referencia.

La *RME* por tanto es una razón de tasas estandarizadas donde la distribución estándar es la de la población expuesta.

Esta estrategia, consistente en utilizar unas “tasas estándar” para obtener la *RME*, suele denominarse *estandarización ‘indirecta’*, aunque el procedimiento de estandarización es idéntico al anterior (también llamado *estandarización ‘directa’*).

Teóricamente, las *RMEs* de dos o más poblaciones no son comparables entre sí debido a que las distribuciones estándar utilizadas en el cálculo de las *RMEs* son diferentes, aunque en muchas situaciones el sesgo que se produce es pequeño. Por otra parte, el error aleatorio es menor que el que se obtendría si se utilizara una distribución estándar diferente de las poblaciones que se comparan (el llamado

“método directo”), por lo que habrá situaciones en las que será preferible utilizar *RMEs*, asumiendo un pequeño sesgo a expensas de reducir el error aleatorio.

Ejemplo 4. En un estudio se analiza la incidencia de cáncer de mama en diversas ocupaciones (4) y se comparan con la población activa de mujeres. En el caso de las médicas se han obtenido los siguientes datos:

Tabla 7. Incidencia de cáncer de mama por ocupación

Edad	Médicas		Población de referencia
	Casos	Personas-año (PT_{1i})	Tasas (I_{Ri})
25-44		11 554	0,00052 año ⁻¹
45-59		12 878	0,00166 año ⁻¹
60-79		4994	0,00238 año ⁻¹
Total	59		

$$w_i = PT_{1i}$$

$$w_{25-44}=11\,554; w_{45-59}=12\,878; w_{60-79}=4994$$

$$RME = \frac{\text{n.º casos observados}}{\text{n.º casos esperados}} = \frac{59}{(0,00052 \times 11\,554) + (0,00166 \times 12\,878) + (0,00238 \times 4994)} = \frac{59}{39,3} = 1,50$$

La incidencia en el grupo de médicas es un 50 % mayor que en la población activa de mujeres, si ambos grupos tuvieran la misma distribución por edad (la de las médicas).

En el caso de las enfermeras y auxiliares la *RME* fue de 0,99 (no se muestran los cálculos). En teoría no se puede realizar una comparación directa con las médicas ya que ambas *RMEs* han sido obtenidas con distribuciones de edad estándar diferentes.

3.1.2. Intervalos de confianza

En la tabla 8 se muestran fórmulas de la varianza de distintos estimadores estandarizados. La fórmula general es la misma que se utiliza en métodos basados en la distribución normal:

$$\text{estimado} \pm Z_{\alpha/2} \sqrt{\text{varianza}(\text{estimado})}$$

En algunos casos se debe usar la escala logarítmica y la fórmula general es

$$\exp \left[\ln(\text{estimado}) \pm Z_{\alpha/2} \sqrt{\text{varianza}(\ln \text{estimado})} \right]$$

Ejemplo 5. Cálculo del intervalo de confianza de las tasas de hipertensos y normotensos del ejemplo 3 (tabla 6). Primero se calculan las varianzas de las tasas específicas por edad:

$$\text{var}(I_{1\ 30-49}) = \frac{3}{2868^2} ; \text{var}(I_{1\ 50-69}) = \frac{80}{13\ 867^2} ; \text{var}(I_{1\ 70+}) = \frac{253}{14\ 842^2}$$

y la varianza de la tasa estandarizada de los hipertensos:

$$\text{var}(I_{1E}) = \frac{\frac{3}{2868^2} \times 31177^2 + \frac{80}{13\ 867^2} \times 44\ 883^2 + \frac{253}{14\ 842^2} \times 30\ 190^2}{(31\ 177 + 44\ 883 + 30\ 190)^2} = 1,98 \times 10^{-7}$$

De forma equivalente se calcula la varianza de la tasa estandarizada de los normotensos

$$\text{Var}(I_{0E}) = 7,21 \times 10^{-8}$$

Y los intervalos de confianza al 90 % de ambas tasas:

$$759 \times 10^{-5} \pm 1,65 \times \sqrt{1,98 \times 10^{-7}} = \mathbf{686 \text{ a } 832} \text{ casos por } 10^5 \text{ personas-año}$$

$$417 \times 10^{-5} \pm 1,65 \times \sqrt{7,21 \times 10^{-8}} = \mathbf{373 \text{ a } 461} \text{ casos por } 10^5 \text{ personas-año}$$

Para el intervalo de confianza al 90 % de la razón de tasas estandarizadas:

$$\text{var}(\ln RT_E) = \frac{1,98 \times 10^{-7}}{(759 \times 10^{-5})^2} + \frac{7,21 \times 10^{-8}}{(417 \times 10^{-5})^2} = 0,00758$$

$$\exp \left[\ln(1,8) \pm 1,65 \times \sqrt{0,00758} \right] = \mathbf{1,56 \text{ a } 2,08}$$

Tabla 8. Fórmulas de la varianza de estimadores estandarizados

Estimador	Varianza
Tasa estandarizada	$\text{var}(I_E) = \frac{\sum \text{var}(I_i) w_i^2}{(\sum w_i)^2}; \quad \text{var}(I_i) = \frac{a_i}{PT_i^2} = \frac{I_i}{PT_i}$
Riesgo estandarizado	$\text{var}(IA_E) = \frac{\sum \text{var}(IA_i) w_i^2}{(\sum w_i)^2}; \quad \text{var}(IA_i) = \frac{IA_i(1 - IA_i)}{N_i - 1}$
Diferencia de tasas estandarizadas	$\text{var}(DT_E) = \text{var}(I_{1E}) + \text{var}(I_{0E})$
Ln de la razón de tasas estandarizadas	$\text{var}(\ln RT_E) = \text{var}(I_{1E})/I_{1E}^2 + \text{var}(I_{0E})/I_{0E}^2$
Diferencia de riesgos estandarizados	$\text{var}(DR_E) = \text{var}(IA_{1E}) + \text{var}(IA_{0E})$
Ln de la razón de riesgos estandarizados	$\text{var}(\ln RR_E) = \text{var}(IA_{1E})/IA_{1E}^2 + \text{var}(IA_{0E})/IA_{0E}^2$

Una fórmula aproximada para calcular el **intervalo de confianza de la RME** es:

$$\exp \left[\ln(RME) \pm Z_{\alpha/2} \frac{1}{\sqrt{O}} \right]$$

Ejemplo 6. Donde O es el número de casos observados. En el ejemplo 4 la RME fue de 1,5. El intervalo de confianza al 90 % es:

$$\exp \left[\ln(1,5) \pm 1,65 \frac{1}{\sqrt{59}} \right] = 1,21 \text{ a } 1,86$$

3.2. Métodos de agregación. Métodos de Woolf y Mantel-Haenszel

Hay un grupo de métodos que utiliza criterios dependientes de los datos del propio estudio para elegir la distribución de pesos que se utilizará para calcular el promedio ponderado de los estimados de los subgrupos. A diferencia de la estandarización, estos métodos se basan en la *asunción de homogeneidad*, por la que se admite que cada subgrupo está estimando el mismo parámetro y que las diferencias observadas entre los subgrupos son debidas a variabilidad aleatoria. Por tanto, una opción lógica para estimar ese valor único es asignar a cada subgrupo un peso inversamente proporcional a su variabilidad. A mayor variabilidad menos peso aporta el subgrupo al promedio. El método de Woolf utiliza como pesos el inverso de la varianza de cada subgrupo y el método Mantel-Haenszel utiliza pesos especiales, proporcionales al inverso de la varianza de cada subgrupo bajo hipótesis nula.

Las variables de estratificación son siempre categóricas; si se quiere estratificar por una variable continua, es necesario agruparla en categorías. Es necesario crear tantos (k) subgrupos como combinaciones de las categorías de las variables de estratificación: si el primer factor de estratificación tuviera, por ejemplo, 3 categorías y el segundo 4, sería necesario crear $k=12$ subgrupos. El análisis estratificado puede sistematizarse en las siguientes etapas:

1. Dividir la muestra en los k subgrupos.
2. Calcular la medida de asociación específica para cada subgrupo i (en el ejemplo utilizaremos la OR , aunque el proceso es análogo para el cálculo de razones y diferencias de tasas, razones y diferencias de riesgos).
3. Comprobar la asunción de homogeneidad, es decir si las OR son similares en todos los subgrupos. Los *test de homogeneidad* ayudan a valorar si las diferencias en las medidas de los subgrupos pueden ser atribuibles a variabilidad aleatoria o no. Si los resultados en los subgrupos son diferentes (es decir, si las variables de estratificación *modifican el efecto* de la variable de exposición principal), se describirán las medidas de asociación en cada subgrupo como resultado final del análisis, en este caso cada una de las OR_i . Si los resultados en los subgrupos son lo suficientemente homogéneos, se entenderá que en todos los subgrupos se está estimando el mismo efecto y que las diferencias encontradas son debidas a variabilidad aleatoria. En este caso, se procede a combinar las estimaciones mediante el cálculo de un promedio ponderado, obteniendo un valor ajustado de la OR , libre de confusión por esa (o esas) variable, ya que se calcula a partir de los estimados específicos de cada subgrupo, que a su vez están libres de confusión por las variables que lo definen. Posteriormente se calcula el intervalo de confianza para la OR combinada.

Supongamos que queremos evaluar el efecto de una variable de exposición principal dicotómica (expuestos y no expuestos) sobre el desarrollo de una enfermedad mediante un diseño de casos y controles, utilizando la odds ratio (OR) como medida de efecto. Se desea valorar el papel de algunas variables que puedan ser potenciales factores de confusión o de modificación de efecto.

Notación para una serie de tablas:

Tabla 9. Notación de los subgrupos en un estudio de casos y controles estratificado

	Expuestos	No expuestos	Total
Casos	a_i	b_i	M_{1i}
Controles	c_i	d_i	M_{0i}
Total	N_{1i}	N_{0i}	T_i

3.2.1. Método de Woolf

La estimación de la OR combinada se calcula como un promedio ponderado, pero en la escala logarítmica, ya que la distribución de la OR es asimétrica (igual que las de la RT y RR). La transformación logarítmica permite utilizar los métodos de la distribución normal. La media ponderada de los logaritmos naturales de los subgrupos es:

$$\ln OR_w = \frac{\sum \ln OR_i w_i}{\sum w_i}$$

Los pesos w_i se calculan como el inverso de la varianza del logaritmo natural de las OR_i

$$w_i = \frac{1}{\text{var}(\ln OR_i)}$$

$$\text{var}(\ln OR_i) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

Ejemplo 7. En el estudio EURAMIC, queremos evaluar la asociación entre tener un índice cintura/cadera, WHR ("waist to hip ratio") por encima de la mediana (expuestos: WHR por encima de la mediana; no expuestos: WHR por debajo de la mediana) y el riesgo de infarto de miocardio (casos y controles), controlando el efecto del consumo de tabaco, agrupado en 3 categorías: nunca fumadores; ex-fumadores y fumadores actuales. La OR cruda es de 3,09.

Tabla 10. Asociación entre índice cintura/cadera, WHR ("waist to hip ratio") y riesgo de infarto de miocardio (casos y controles), por consumo de tabaco

	WHR alto	WHR bajo	OR_i	$\ln OR_i$	$\text{var}(\ln OR_i)$	$w_i = 1/\text{var}(\ln OR_i)$
No fumadores						
Casos	39	17				
Controles	65	86	3,04	1,11	0,111	9,0
Exfumadores						
Casos	108	27				
Controles	110	102	3,71	1,31	0,065	15,3
Fumadores						
Casos	239	80				
Controles	126	111	2,63	0,97	0,034	29,7

Para valorar la homogeneidad de k subgrupos, se puede utilizar un test de χ^2 con $k - 1$ grados de libertad, contrastando la hipótesis nula de que $OR_1 = OR_2 = \dots = OR_k$

$$\chi_{k-1}^2 = \sum \frac{(\ln OR_i - \ln OR_w)^2}{\text{var}(\ln OR_i)}$$

Siguiendo el **ejemplo 7**:

$$\ln OR_w = \frac{\sum \ln OR_i w_i}{\sum w_i} = \frac{1,11 \times 9,0 + 1,31 \times 15,3 + 0,97 \times 29,7}{9,0 + 15,3 + 29,7} = 1,09$$

$$\chi_{k-1}^2 = \frac{(1,11 - 1,09)^2}{0,111} + \frac{(1,31 - 1,09)^2}{0,065} + \frac{(0,97 - 1,09)^2}{0,034} = 1,2$$

Con este valor de χ^2 se obtiene el valor P asociado a la hipótesis nula. Si este valor P es mayor de un nivel predeterminado, por ejemplo de 0,15, se asume que los datos han superado el valor mínimo de consistencia con la homogeneidad que se consideró aceptable y, por tanto, se puede calcular el estimado puntual del efecto agrupado OR_w y su intervalo de confianza.

En el **ejemplo 7** el valor P correspondiente a una χ^2 con 2 grados de libertad de 1,2 es 0,55 por lo que supera el valor predeterminado y por tanto se decide que las diferencias entre las OR_i son compatibles con variabilidad aleatoria, y se procede a presentar el promedio ponderado OR_w y su intervalo de confianza.

$$OR_w = \exp(\ln OR_w)$$

$$OR_w = \exp(1,09) = 2,97$$

El intervalo de confianza de la OR_w :

$$\exp[\ln OR_w \pm Z_{\alpha/2} EE(\ln OR_w)]$$

El error estándar de $\ln OR_w$ es:

$$EE(\ln OR_w) = \sqrt{\frac{1}{\sum w_i}}$$

En el **ejemplo 7**:

$$EE(\ln OR_w) = \sqrt{\frac{1}{54}} = 0,136$$

Y el intervalo de confianza al 90 %:

$$\exp[1,09 \pm 1,65 \times 0,136] = 2,38 \text{ a } 3,72$$

Tomando la OR como una buena estimación de la razón de riesgos, se puede interpretar como que aquellas personas con una razón cintura/cadera por encima de la mediana tienen un riesgo de infarto que casi triplica el de las personas con razón cintura cadera menor que la mediana, a igualdad de categoría de consumo de tabaco. No se aprecia confusión por esta variable ya que la OR cruda, 3,09, es prácticamente igual a la ajustada, 2,97.

3.2.2. Método de Mantel-Haenszel

Mantel y Haenszel propusieron una fórmula sencilla de calcular la OR ajustada. Entre los métodos de agregación se considera el mejor en la mayoría de las situaciones:

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{T_i}}{\sum \frac{b_i c_i}{T_i}}$$

El error estándar del $\ln OR_{MH}$ (Robins *et al* (5,6)):

$$EE(\ln OR_{MH}) = \sqrt{\frac{\sum P_i R_i}{2(\sum R_i)^2} + \frac{\sum (P_i S_i + Q_i R_i)}{2(\sum R_i)(\sum S_i)} + \frac{\sum Q_i S_i}{2(\sum S_i)^2}}$$

donde

$$P_i = \frac{a_i + d_i}{T_i}; Q_i = \frac{b_i + c_i}{T_i}; R_i = \frac{a_i d_i}{T_i}; S_i = \frac{b_i c_i}{T_i}$$

Así, el *intervalo de confianza* para OR_{MH} :

$$\exp[\ln OR_{MH} \pm Z_{\alpha/2} EE(\ln OR_{MH})]$$

Ejemplo 8. La estimación de la OR combinada por el método de Mantel-Haenszel con los datos del ejemplo 7:

$$OR_{MH} = \frac{\frac{39 \cdot 86}{207} + \frac{108 \cdot 102}{347} + \frac{239 \cdot 111}{556}}{\frac{17 \cdot 65}{207} + \frac{27 \cdot 110}{347} + \frac{80 \cdot 126}{556}} = 2,99$$

que es similar a la obtenida por el método de Woolf ($OR_w = 2,97$).

El error estándar del $\ln OR_{MH}$: 0,136, con el que podemos calcular intervalos de confianza. Por ejemplo, el intervalo de confianza al 90 % es:

$$\exp[\ln 2,99 \pm 1,65 \times 0,136] = 2,39 \text{ a } 3,74$$

3.2.3. Estimadores de Mantel-Haenszel en otros diseños

Existen fórmulas equivalentes para otros estimadores combinados de Mantel-Haenszel (diferencia de tasas, razón de tasas, diferencia de riesgos, razón de riesgos) y sus varianzas. Se recomienda consultar estas fórmulas en el libro de Rothman, Greenland y Lash. (7)

3.3. Diferencias entre los métodos de análisis estratificado

Los métodos de cálculo de estimadores ajustados por *agregación*, como los de Woolf y Mantel-Haenszel, asumen homogeneidad del efecto en los subgrupos. La estandarización sin embargo no asume homogeneidad del efecto entre subgrupos, y la elección de los pesos puede realizarse con cualquier criterio, que conviene justificar. Si se aprecia modificación del efecto, lo esencial es presentar e interpretar los estimados de los subgrupos, pero si además se desea obtener una medida global del efecto neto de la exposición es preferible utilizar la estandarización, ya que los pesos de Woolf y Mantel-Haenszel no reflejan una población diana reconocible. En los diseños de casos y controles también se puede utilizar estandarización. (7)

En cuanto a las diferencias entre el método de Woolf y el de Mantel-Haenszel, cabe resaltar que el método de Woolf requiere que todas las celdas de cada subgrupo incluyan alto número de efectivos; si no es así, como ocurre en la mayoría de las ocasiones, es muy probable que las estimaciones sean erróneas. Esta limitación es el principal problema del análisis estratificado: cuando el número de factores de confusión es alto, los subgrupos van teniendo menos efectivos y las celdas van quedando vacías.

Supongamos, en una situación muy común, que disponemos de un tamaño muestral de 800 personas en un estudio de casos y controles, y que deseamos controlar por edad en 3 categorías, por sexo en 2, por clase social en 3, y por consumo de tabaco en 3. El número de subgrupos es de: $3 \times 2 \times 3 \times 3 = 54$. En el caso (por otra parte muy raro) de que las 800 personas se distribuyeran equitativamente en los subgrupos, éstos tendrían de promedio 14,8 personas. Lo más habitual es que hubiera algunos subgrupos sin casos o sin controles. Si además consideramos conveniente ajustar por hipertensión en dos categorías, el número total será de: $54 \times 2 = 108$, reduciéndose el promedio a 7,4. Este problema es menor si se utiliza el método de Mantel-Haenszel, que permite obtener estimaciones válidas con muy pocos efectivos en cada subgrupo, incluso con algunas celdas vacías, cosa que no es posible con el método de Woolf. Esta y otras propiedades hacen de éste el *método de elección* (entre los métodos que asumen homogeneidad), añadiendo a sus ventajas la facilidad de cálculo. Aun así, en muchas situaciones el número de subgrupos en relación con el tamaño de la muestra será excesivo y se tiene que re-

currir a la utilización de modelos de regresión. En cualquier caso, incluso cuando finalmente se utilicen modelos de regresión, es conveniente realizar análisis estratificado, incluyendo aquellas variables con mayor relevancia, bien científica o de salud pública, ya que este análisis permitirá “conocer” mejor los datos, algo que es más complicado en el caso de usar modelos de regresión multivariantes.

3.4. Confusión residual

Se habla de confusión residual cuando, por alguna razón, no se ha controlado toda la potencial confusión presente en los datos recogidos en un estudio. Esta situación puede darse cuando no se ha ajustado por todas las variables de confusión o porque al categorizar variables de confusión continuas en un número pequeño de subgrupos no se puede asumir ausencia de confusión dentro de cada subgrupo. Por ejemplo, si se desea controlar el efecto de la edad en una población entre 15 y 65 años y se crean sólo dos subgrupos (p. ej.: 15-35 años y 36-65), puede haber diferencias sustanciales de edad entre las poblaciones que se comparan dentro de cada subgrupo y por tanto haber confusión (el caso extremo es *un* solo “subgrupo”, es decir datos crudos). Normalmente, con pocos subgrupos se controla bien la confusión en la mayoría de las situaciones, siendo la estrategia más adecuada comprobar cuál es el número mínimo de subgrupos que garantice un nivel de control que se considere adecuado.

Por último, habrá confusión residual si las variables de confusión tienen gran error de medida.

3.5. Aplicación al análisis de datos emparejados

En los estudios de casos y controles emparejados suele ser necesario controlar por las variables de emparejamiento ya que si no se introducirá un sesgo (en general hacia el valor nulo). En estos estudios, el análisis se realiza estratificando por parejas; es decir cada pareja caso-control forma un subgrupo. El análisis entonces se realiza mediante la fórmula de Mantel-Haenszel, apropiada para datos esparcidos (pocas observaciones en muchos subgrupos). Los subgrupos pueden clasificarse en 4 tipos: dos en los que la pareja es concordante (R y V) en relación con la exposición y dos tipos en los que son discordantes (S y U):

Tabla 11. Distribución tipos de parejas en un estudio de casos y controles con emparejamiento

	R		S		U		V	
	caso	control	caso	control	caso	control	caso	control
Exp	1	1	1	0	0	1	0	0
No exp	0	0	0	1	1	0	1	1

$$\frac{a_i d_i}{T_i} = \frac{b_i c_i}{T_i} = 0 \quad \frac{a_i d_i}{T_i} = \frac{1}{2}; \frac{b_i c_i}{T_i} = 0 \quad \frac{b_i c_i}{T_i} = \frac{1}{2}; \frac{a_i d_i}{T_i} = 0 \quad \frac{a_i d_i}{T_i} = \frac{b_i c_i}{T_i} = 0$$

Como se puede apreciar, las parejas concordantes (R y V) no aportan información a la OR combinada por lo que la fórmula se reduce a la razón entre el número de parejas discordantes en las que el caso está expuesto (S) y en las que lo está el control (U):

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{T_i}}{\sum \frac{b_i c_i}{T_i}}; \quad OR_{MH} = \frac{S(1/2)}{U(1/2)} = \frac{S}{U}$$

Para calcular los intervalos de confianza de la OR_{MH} se necesita calcular el error estándar del logaritmo natural de la OR_{MH} , que en el caso de datos emparejados se simplifica a:

$$EE(\ln OR_{MH}) = \sqrt{\frac{1}{S} + \frac{1}{U}}$$

Y los límites del intervalo de confianza de la OR_{MH} :

$$\exp[\ln OR_{MH} \pm Z_{\alpha/2} EE(\ln OR_{MH})]$$

Ejemplo 9. Davey et al (8) usaron la base de datos de servicios de medicina general (general practices, que prestan los servicios de atención primaria en el Sistema Nacional Salud) del Reino Unido para estudiar el posible efecto de la migraña en la ocurrencia de asma. Seleccionaron 64 678 casos que se emparejaron con un número igual de controles del mismo sexo, año de nacimiento y servicio de atención primaria. El número (S) de parejas, en las que el caso de asma tenía antecedentes de migraña y su control no, fue de 7300. El número (U) de parejas en las que el caso de asma no tenía antecedentes de migraña y su control sí, fue de 4584. La OR se calcula como la razón entre estos números de parejas discordantes:

$$OR_{MH} = \frac{S}{U} = \frac{7300}{4584} = 1,59$$

$$EE(\ln OR_{MH}) = \sqrt{\frac{1}{7300} + \frac{1}{4584}} = 0,0188$$

Y el intervalo de confianza al 99%:

$$\exp[\ln 1,59 \pm 2,58 \times 0,0188] = 1,52 \text{ a } 1,67$$

4. Referencias

1. Greenland S, Senn SJ, Rothman KJ, *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol.* 2016;31(4):337-350. doi:[10.1007/s10654-016-0149-3](https://doi.org/10.1007/s10654-016-0149-3).
2. Miettinen OS. *Theoretical Epidemiology*. John Wiley & Sons; 1985.
3. Poole C. Beyond the confidence interval. *Am J Public Health.* 1987;77(2):195-199.
4. Pollán M, Gustavsson P. High-risk occupations for breast cancer in the Swedish female working population. *Am J Public Health.* 1999;89(6):875-881. doi:[10.2105/ajph.89.6.875](https://doi.org/10.2105/ajph.89.6.875).
5. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am J Epidemiol.* 1986;124(5):719-723. doi:[10.1093/oxfordjournals.aje.a114447](https://doi.org/10.1093/oxfordjournals.aje.a114447).
6. Robins J, Breslow N, Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics.* 1986;42(2):311-323.
7. Greenland S, Rothman KJ. Introduction to stratified analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Lippincott William & Wilkins; 2008:258-282.
8. Davey G, Sedgwick P, Maier W, Visick G, Strachan DP, Anderson HR. Association between migraine and asthma: matched case-control study. *Br J Gen Pract.* 2002;52(482):723-727.

Capítulo 7.

Gráficos causales

MIGUEL ÁNGEL ROYO BORDONADA y JOSÉ RAMÓN BANEGAS BANEGAS

1. Introducción

2. Gráficos causales (dirigidos y no cíclicos)

3. Representación de los sesgos

3.1. Sesgos de selección

3.2. Sesgos de confusión

3.3. Sesgos de información

4. Control de la confusión

4.1. Conjunto mínimo suficiente de variables

4.2. Ventajas y limitaciones de los gráficos causales

4.3. Principios para el control de la confusión

5. Referencias

1. Introducción

En un estudio epidemiológico, el análisis de la hipótesis de trabajo sobre una relación causal entre la exposición (*exposure*, en inglés) y el desenlace (*outcome*) de interés requiere la consideración de terceras variables, los denominados factores de confusión o confundidores, comúnmente denominados confusores (*confounders*), que podrían sesgar los resultados (*confounding bias*). Los criterios convencionales, considerados necesarios, pero insuficientes, para que una variable sea un factor de confusión son los siguientes (1):

1. La variable está asociada con la exposición en la población.
2. La variable es una causa o marcador de una causa del desenlace. Es decir, la variable está asociada con el desenlace también en la población no expuesta.
3. La variable no es un mediador en la relación o cadena causal entre la exposición y el desenlace.

Como veremos más adelante, estos criterios pueden resultar insuficientes o inadecuados para decidir las variables por las que es necesario ajustar en los análisis de una hipotética relación causal entre una exposición y un desenlace cuando se consideran múltiples factores de confusión en los estudios epidemiológicos observacionales. (2) Para decidir qué variables hay que controlar se ha recurrido tradicionalmente a criterios estadísticos en los modelos de regresión multivariable, mediante procedimientos de selección paso a paso, incorporando todas las variables que producen una mejora en el ajuste estadístico del modelo, de forma que no dejemos de controlar ningún factor de confusión, aun a costa de incluir en el modelo variables no confusoras, o seleccionando solo aquellas variables que producen tras su ajuste un cambio en el estimador del efecto de al menos el 10 %, asumiendo que así controlamos los factores de confusión relevantes. (3) Otros criterios aplicados tradicionalmente consisten en ajustar por (*adjust for*) todas las variables medidas previamente a la exposición de interés o por todas las variables que cumplen los criterios convencionales para ser consideradas factores de confusión. Cualquiera de estas estrategias puede producir un resultado sesgado al ajustar por variables que no son factores de confusión (sobreajuste) o por omisión de algún factor de confusión (infra-ajuste o confusión residual). Para evitar este problema, es esencial que la selección de las variables a controlar se base en el conocimiento previo (biológico, psico-social, etc.) sobre las relaciones existentes entre terceras variables y la exposición o el desenlace. Los gráficos causales sirven para representar

esa red de relaciones («estructura causal» de la hipótesis en estudio) y, en función de la misma, identificar sesgos de confusión, selección e información, y ayudar a seleccionar las variables que hay que controlar en el análisis de los datos.

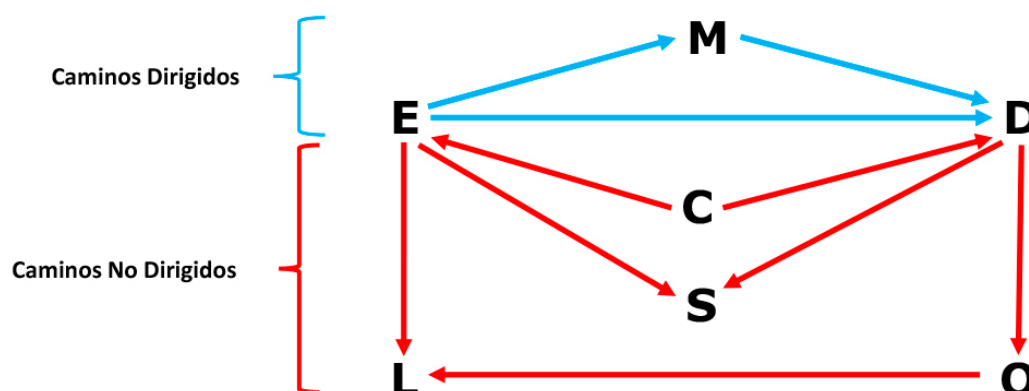
Utilidad de los gráficos causales

1. Representar la red o estructura de relaciones causales de la hipótesis en estudio.
2. Identificar, en dicha red, sesgos de confusión, selección e información.
3. Seleccionar las variables por las que es necesario ajustar en el análisis de los datos e identificar aquellas por las que no se debe ajustar.

2. Gráficos causales (dirigidos y no cíclicos)

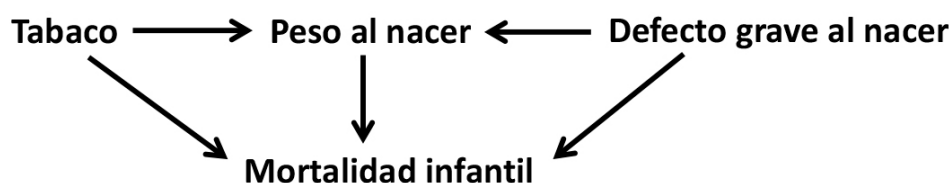
Un gráfico causal es un diagrama compuesto por nodos, que representan variables, y flechas, que representan relaciones causales directas entre variables adyacentes, con el astil de la flecha partiendo de la causa y la punta dirigiéndose al desenlace o efecto: por ejemplo, en la figura 1, E es una causa directa de D, M, S, y L, O es una causa directa de L, M es una causa directa de D, y C es una causa directa de E y D. (4) La ausencia de una flecha entre dos variables cualesquiera implica que no existe relación causal directa entre esas dos variables. El gráfico causal representa las asunciones del investigador, basadas en el conocimiento disponible sobre las relaciones causales entre las variables contenidas en el mismo. Como las causas preceden a sus efectos, **los gráficos causales no son cíclicos**: uno nunca puede comenzar desde una variable y, siguiendo la dirección de las flechas, terminar en la misma variable: el consumo de tabaco causa cáncer de pulmón, pero el cáncer de pulmón no puede afectar al hábito tabáquico previo a la fecha de su aparición.

Figura 1. Caminos dirigidos y no dirigidos en un gráfico causal (E: exposición; D: desenlace; M y O: mediadores; C: confusor, causa común de E y D; S y L: colisionadores, efectos comunes de E y D). (4)



En un gráfico causal, un camino (*path*) entre dos nodos es una ruta trazada a lo largo de una o varias flechas que conectan nodos adyacentes. Los caminos pueden recorrerse en los dos sentidos, excepto cuando nos encontramos dos flechas apuntando al mismo nodo, lo que se denomina un factor de colisión o colisionador (*collider*), en cuyo caso decimos que el camino está bloqueado o cerrado. Por ejemplo, en el gráfico causal de la figura 1, hay cinco caminos que conectan E con D, tres abiertos (E-D, E-M-D y E-C-D) y dos cerrados (E-S-D, bloqueado en S, y E-L-O-D, bloqueado en L). En otras palabras, E y D colisionan (convergen) en S, y E y O colisionan en L. Ejemplo: la llamada paradoja del peso al nacer se derivaba del hecho de que los niños de madres fumadoras tenían un peso al nacer más bajo que los de madres no fumadoras, pero morían menos. (5) La paradoja se resolvió al analizar los datos con la ayuda de gráficos causales, donde se identificó que el peso al nacer actuaba como factor de colisión del tabaquismo materno y los defectos graves al nacer (figura 2). Para entender cómo se produce esta paradoja con la ayuda de los gráficos causales, es necesario conocer antes la distinción entre caminos dirigidos y no dirigidos y entre caminos no dirigidos abiertos y cerrados.

Figura 2. Paradoja del peso al nacer (Exposición: tabaquismo materno; Desenlace: mortalidad del niño; Confusor: defecto al nacer; Colisionador: peso al nacer). (5)



Un **camino dirigido** es aquel que se recorre siempre en el sentido de las flechas; en la figura 1 E-D y E-M-D son caminos dirigidos, uno de ellos directo (E-D) y el otro indirecto (E-M-D); mientras que el resto son caminos no dirigidos. **Los gráficos causales son dirigidos** (*directed acyclic graphs: DAGs*): una variable es una causa de otra variable si existe un camino dirigido que partiendo de la primera conduce a la segunda, y se trata de una causa directa cuando no existe un nodo intermedio entre las dos variables. El concepto de causa directa tiene que ver con los límites del conocimiento disponible acerca de los mecanismos causales o con la ausencia de representación de los mismos en el gráfico causal. **Ejemplo:** Desde los años 50 del siglo XX, a raíz de dos estudios epidemiológicos de Doll y de Hammond, (6,7) se sabe que el tabaco es una causa de cáncer de pulmón; sin embargo, tuvieron que pasar varias décadas hasta que se conocieron algunos de los mecanismos moleculares implicados en el efecto del tabaco, (8) un conocimiento que se sigue ampliando en los últimos años. (9) Así, un gráfico causal con una sola flecha

entre las variables tabaco y cáncer de pulmón representa una relación causal en la que no se conocen o no se han representado los mecanismos a través de los cuales el tabaco causa cáncer de pulmón. Pueden existir varios mecanismos mediadores entre una causa y su efecto, algunos conocidos y otros no, lo que se representa en el gráfico causal mediante varios caminos dirigidos entre ambas variables, entre ellos uno directo (mediadores desconocidos, ocasionalmente inexistentes o muy próximos en el tiempo); por ejemplo, en la figura 1, E es causa de D, en parte, a través de M (variable mediadora) y, en parte, de forma directa, donde la flecha que va directamente de E a D representa el efecto causal de E sobre D no debido a M.

Un **camino no dirigido** es aquel que, en parte, se recorre en el sentido opuesto al de las flechas. Un camino no dirigido que parta de la exposición en el sentido contrario al de la flecha (i.e., la exposición «recibe» una flecha) es un «camino por la puerta de atrás» (*backdoor path*). En otras palabras, la puerta de atrás es una vía abierta o paso alternativo (al causal) y una manera gráfica de denominar a un efecto de confusión en los gráficos causales. En la figura 1, de los tres caminos no dirigidos entre E y D, únicamente E-C-D es un camino por la puerta de atrás. Un camino no dirigido que parta de la exposición en el sentido de la flecha (i.e., E «lanza» una flecha) es un camino cerrado o bloqueado. En la figura 1, E-L-O-D y E-S-D son dos caminos cerrados entre E y D, ya que están bloqueados en L y S respectivamente. En el ejemplo de la figura 2, al examinar solo bebés con bajo peso al nacer, estamos condicionando por ese colisionador. Esto abre un camino no dirigido que parte de la exposición en el sentido de la flecha, creando así una asociación espuria entre el tabaco y la mortalidad infantil, que sigue los siguientes pasos: 1) de tabaco a peso al nacer, 2) de peso al nacer a defecto grave al nacer y 3) de defecto grave al nacer a mortalidad infantil. El segundo paso no es causal, porque la flecha va en sentido inverso, es decir, del defecto grave al nacer al peso al nacer; pero induce una asociación sesgada entre tabaco y mortalidad infantil, de tal modo que la relación parece protectora, dado que, entre los niños con bajo peso al nacer, la frecuencia de defectos graves al nacer es menor en los hijos de madres fumadoras.

La tabla 1 presenta un resumen de las diferencias entre confusores y colisionadores, que se ilustran con otro ejemplo en el siguiente apartado.

Tabla 1. Diferencias entre confusor y colisionador

	Confusor	Colisionador
Característica	Causa común	Efecto común
Asociación	Contribuye a la asociación entre sus efectos	No contribuye a la asociación entre sus causas
Tipo de paso	Paso abierto	Paso cerrado
Efecto tras ajustar (condicionar)	Paso cerrado	Paso abierto
¿Sesgo antes de ajustar?	Sí, sesgo de confusión	No
¿Sesgo tras ajustar?	No	Sí, sesgo de colisión

Fuente: Elaboración propia (adaptada de South Australia Univesity: *Introduction to confounding and Directed Acyclic Graphs. Risk and Casualty. Online Modules in Research Methods and Data Analysis*).

3. Representación de los sesgos

3.1. Sesgos de selección

Los sesgos de selección en los estudios epidemiológicos se dan cuando el valor del parámetro de interés en la población objetivo es diferente del valor del parámetro en la muestra de sujetos disponible en el estudio. Este sesgo se produce cuando la probabilidad de participar en el estudio es distinta para los sujetos de la población objetivo en función de que pertenezcan a una u otra categoría de las variables de exposición o desenlace; habitualmente, por seleccionar a los participantes con base en la ausencia o presencia (en mayor o menor grado) de una (o varias) tercera variable —una forma de condicionar por esa variable (en términos epidemiológicos, condicionar es lo mismo que ajustar o controlar)— que se asocia con las variables exposición y desenlace creando una asociación espuria entre estas.

Ejemplo: la obesidad se asocia con una menor mortalidad en pacientes con enfermedades crónicas, lo que se conoce como la «paradoja de la obesidad». (10) Así, entre los pacientes con diabetes tipo 2 y comorbilidad cardiovascular, los que presentan obesidad tienen menor mortalidad que los que presentan un peso normal. (11) Entre los pacientes con diabetes seleccionados para participar en un estudio epidemiológico, los que no presentan obesidad han desarrollado diabetes por otras causas (no medidas): podrían ser genéticas, estilos de vida, exposiciones ambientales o cualquier

combinación de las mismas. Pues bien, como estas otras causas de diabetes también aumentan la mortalidad por mecanismos diferentes a la diabetes, en diabéticos se da una asociación inversa, de carácter espurio, entre obesidad y mortalidad que explica la paradoja de la obesidad. (12) Para entender cómo se produce el sesgo, volveremos a este ejemplo, ampliado y graficado (figuras 3.a y 3.b), en el siguiente párrafo; pero antes conviene explicar la representación gráfica de la independencia estadística.

Figura 3.a. Paradoja de la obesidad. (13)

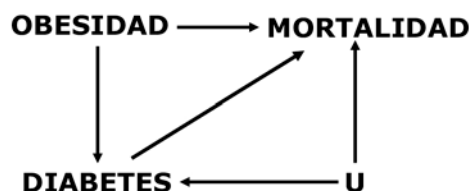
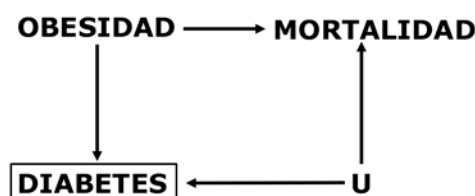


Figura 3.b. Paradoja de la obesidad. (13)



La independencia (ausencia de asociación marginal) entre dos variables se representa gráficamente por la ausencia de caminos abiertos entre las mismas. Resulta evidente que un mismo efecto causado por dos variables sin que una sea mediadora de la otra (factor de colisión) no puede inducir una asociación entre esas variables, ya que las causas preceden al efecto y, por tanto, la relación (o no) entre las variables causantes ya estaba establecida cuando se produjo el efecto y no puede ser modificada por este. También cabe la posibilidad de que no exista asociación marginal entre dos variables en presencia de varios caminos abiertos entre las mismas, cuando las asociaciones debidas a los mismos se cancelan entre sí, aunque esto es raro de observar en la práctica. (2) En la figura 3.a se muestra una red (simplificada a efectos didácticos) de variables en una población de personas con y sin diabetes, donde la variable U representa las causas no medidas de la diabetes (*unobserved or unmesaruded variables*). (12) Pues bien, en este gráfico hay tres caminos entre la exposición (obesidad) y el desenlace (mortalidad), dos dirigidos (Obesidad-Mortalidad y Obesidad-Diabetes-Mortalidad), y por tanto causales, y uno no dirigido (Obesidad-Diabetes-U-Mortalidad), que está bloqueado en Diabetes (factor de colisión). La relación entre Obesidad y U no puede verse afectada por Diabetes, ya que esta se produce con posterioridad a la presencia de Obesidad y U. Por tanto, Obesidad y U no están asociadas, ya que el único cami-

no entre las mismas está cerrado, y, por ende, tampoco existe una asociación entre Obesidad y Mortalidad por ese camino. Por el contrario, en el gráfico de la figura 3.b se muestra la misma red variables en una población (seleccionada) de personas con diabetes, que se representa gráficamente enmarcando la variable por la que se seleccionan los participantes (diabetes). Al seleccionar personas con diabetes, que a efectos estadísticos es equivalente a ajustar por esa variable (método de restricción), hemos controlado el efecto mediador de la diabetes en la relación entre obesidad y mortalidad; por tanto, en esta población el efecto de la obesidad en la mortalidad es únicamente aquella parte del efecto global que no está mediada por la diabetes, representada por el camino dirigido directo entre ambas variables. Pero, al mismo tiempo, hemos creado una asociación espuria, meramente estadística, entre obesidad y mortalidad, ya que las causas no medidas de la diabetes (U) también aumentan la mortalidad por mecanismos ajenos a la propia diabetes. Por tanto, al seleccionar personas con diabetes y analizar la relación entre obesidad y mortalidad, que es una forma de condicionar por esa variable (factor de colisión), hemos abierto un camino no dirigido entre obesidad y mortalidad, creando una asociación negativa, de carácter espurio, entre ambas variables. La asociación neta será el resultado de la combinación entre el efecto meramente estadístico del sesgo de selección y el efecto real de la obesidad en la mortalidad (no mediado por la diabetes), que dará lugar a una atenuación o inversión del efecto observado en función de la magnitud de ambas asociaciones. Asimismo, en un estudio en el que participasen personas con y sin diabetes, incluso aunque participase toda la población objetivo, también se podría abrir un camino no dirigido condicionando por la variable diabetes (factor de colisión) mediante técnicas de ajuste estadístico (estratificación o regresión multivariable). Es decir, controlar por una variable que es un factor de colisión en el camino entre la exposición y la enfermedad, tanto si se produce por selección de los sujetos que participan en el estudio como condicionando los análisis por esa variable (ajuste indebido) suele introducir un sesgo denominado sesgo de colisión. (13) El sesgo de colisión es, por tanto, un concepto más amplio que el de sesgo de selección, ya que abarca la mayoría de los sesgos de selección (pero no todos, como veremos a continuación) y los sesgos de confusión por ajuste estadístico indebido por variables que son factores de colisión.

Utilidad de la representación gráfica de la colisión

- Detección de la presencia de sesgos de selección: cuando las variables que han influido en la selección de los participantes podrían ser un efecto común de la exposición y el desenlace.
- Visualización de las posibilidades de prevención de los sesgos de colisión:
 - No restringir la población de estudio por factores de colisión.
 - No ajustar en los análisis estadísticos por factores de colisión.

Los sesgos de selección también pueden producirse en ausencia de un factor de colisión. En un ejemplo hipotético de un estudio de cohortes sobre la relación causal entre E y D, con un $RR=1,69$ en la población objetivo y un 28 % de sujetos censurados (sin información de la variable resultado) por rechazo a participar o pérdida del seguimiento, el RR en los participantes no censurados fue de 2,05. (14) En este ejemplo, el porcentaje de sujetos censurados (28 %) fue el mismo en expuestos y no expuestos. Por tanto, Greenland demostró que es suficiente con que la censura se asocie con el desenlace, aunque no lo haga con la exposición, para que se produzca sesgo de selección. Hernán representó este ejemplo mediante el gráfico causal de la figura 4 en los sujetos no censurados, donde «U» representa una variable (o combinación de variables) no medidas con un efecto causal sobre el desenlace y la probabilidad de ser censurado. (15) Aplicando las reglas de la «separación gráfica» que veremos más adelante, se demuestra que E y D están asociados en la población objetivo y en la muestra de sujetos no censurados, y puesto que el valor del RR es diferente en ambos grupos, estamos ante un sesgo de selección en ausencia de un factor de colisión. Es decir, bajo la hipótesis alternativa (presencia de asociación en la población objetivo), condicionar en K, que no es un factor de colisión, induce sesgo de selección. Por el contrario, aplicando las mismas reglas en la figura 5, bajo la hipótesis nula (E y D no asociadas en la población objetivo) condicionar en K no induce sesgo de selección. El sesgo de selección bajo la hipótesis alternativa requiere que la asociación entre K y D varíe a través de los niveles de E en el parámetro usado para medir el efecto de E en D (razón o diferencia de riesgos) en la población objetivo. La incapacidad de los gráficos causales para codificar el sesgo de selección en ausencia de un factor de colisión se debe a la dependencia de este sesgo de una particular parametrización del efecto, ya que los gráficos causales no son paramétricos. En resumen, condicionar en un factor de colisión inducirá una asociación espuria entre la exposición y el resultado, incluso bajo la hipótesis nula (ausencia de relación causal en la población objetivo), razón por la que algunos epidemiólogos se refieren a este sesgo como sesgo de selección bajo la hipótesis nula (*selection bias under the null*); mientras que condicionar en una variable que no es un factor de colisión puede a lo sumo inducir una asociación espuria bajo la hipótesis alternativa (presencia de relación causal en la población objetivo), que sesgará el efecto observado. (15)

Figura 4. Gráfico causal bajo la hipótesis alternativa en población no censurada (E: exposición, D: desenlace, K: censura; U: variable o variables no medidas). (15)

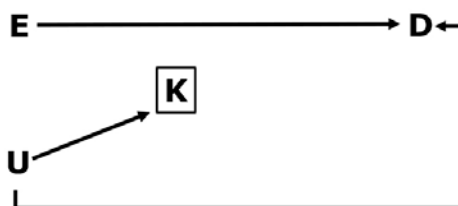
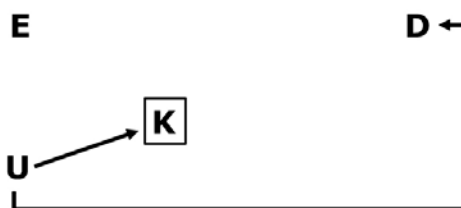


Figura 5. Gráfico causal bajo la hipótesis nula en población no censurada (E: exposición, D: desenlace, K: censura; U: variable o variables no medidas). (15)



3.2. Sesgos de confusión

Los factores de confusión se representan gráficamente como lo que son: causas comunes de la exposición y el desenlace. Por ejemplo, en el gráfico de la figura 1, la relación entre E y D está confundida por C. En general, la asociación entre dos variables se representa gráficamente por la existencia de al menos un camino abierto entre las mismas, que será causal cuando el camino es dirigido o una mera asociación estadística cuando el camino es no dirigido, como en los sesgos de selección o como en este caso, a través de una causa común (factor de confusión), en cuyo caso decimos que la asociación está confundida, ya que los grupos de expuestos y no expuestos no son «intercambiables» (*nonexchangeable*); es decir, no son comparables bajo un modelo contrafáctico (ver capítulo de estudios de intervención). Como hemos visto en capítulos previos, podemos controlar la confusión condicionando por las variables confusoras, bien restringiendo el análisis a los participantes en función de su exposición a las mismas (seleccionar únicamente a participantes expuestos —o no expuestos— a los factores de confusión), bien ajustando estadísticamente por las variables de confusión mediante estratificación o regresión multivariable. Pero, como hemos mostrado previamente, si ajustamos en los análisis por un factor de colisión, al abrir un camino no dirigido entre E y D, estamos creando una asociación espuria, meramente estadística, entre ambas variables. Si bien hemos caracterizado ese error como sesgo de selección, a efectos prácticos el resultado es el mismo que ante la existencia de un factor de confusión, en que la asociación entre E y D está confundida, porque los grupos de expuestos y no expuestos en las categorías del factor de colisión no son intercambiables.

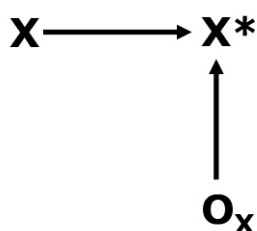
Representación gráfica de la confusión

- En los gráficos causales, los factores de confusión se definen como causas comunes de la exposición y el desenlace; mientras que la confusión se define como un sesgo debido a la existencia de un camino abierto por la puerta de atrás entre exposición y desenlace.
- Estas definiciones de confusión y factores de confusión no se basan en las asociaciones estadísticas de los datos de un estudio, sino en conocimientos cualitativos previos sobre la estructura causal del problema, que codificamos gráficamente.

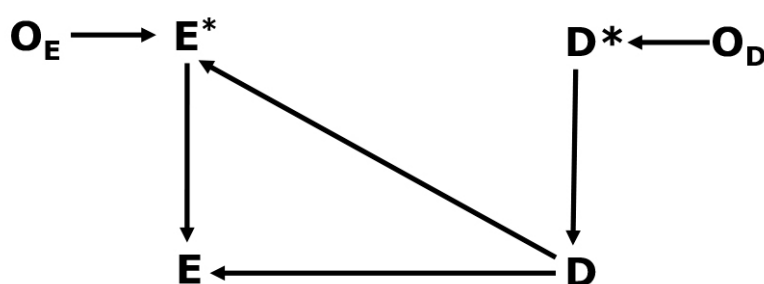
3.3. Sesgos de información

Los sesgos de información son errores sistemáticos que se dan cuando el valor del parámetro de interés en una población determinada difiere de su valor verdadero por errores en la medición de las variables. Cuando se trata de variables discretas, este sesgo es conocido como error de clasificación. Dada una variable X , denominaremos X^* a esa misma variable cuando exista un error de medición. El valor de X^* depende en parte de X , lo que se representa gráficamente como una de sus causas directas (figura 6), y en parte de otra variable o conjunto de variables, conocidas o no, que representamos como otra de sus causas directas (O_x). Los sesgos de información también pueden ocurrir en estudios de intervención aleatorizados, dando lugar a un error en el estimador de la asociación entre exposición y desenlace incluso en un estudio de intervención ideal, en el que esté garantizada la intercambiabilidad de los grupos de expuestos y no expuestos.

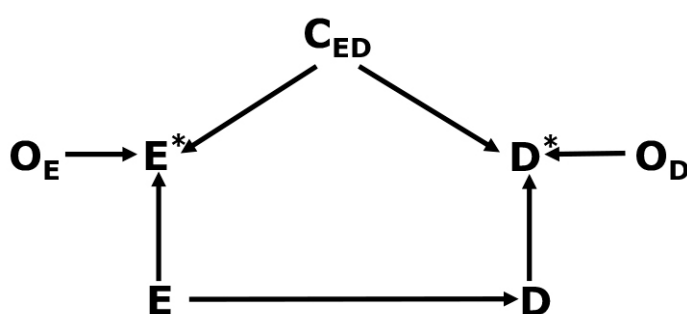
Figura 6. Sesgo de información. (16)



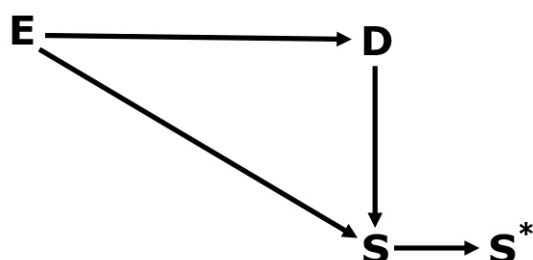
La estructura causal de la confusión se representa gráficamente como la presencia de «causas comunes» de la exposición y el desenlace, y la estructura causal del sesgo de selección como el condicionamiento sobre «efectos comunes» de la exposición y el desenlace (o sus causas), pero no hay una representación única para la estructura causal de los sesgos de información, ya que esta depende, por un lado, de si el sesgo de información es diferencial o no diferencial, y, por otro lado, en presencia de errores de medición tanto en la exposición como el desenlace, de si estos dos errores son independientes o no. (16)

Figura 7. Sesgo de información diferencial con errores de medición independientes. (16)

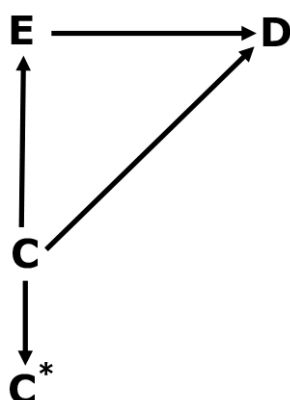
Cuando el error en la medición de la exposición no depende del desenlace o cuando el error en la medición del desenlace no depende de la exposición, hablamos de errores de medición no diferenciales. Por el contrario, cuando el error de medición de la exposición depende del desenlace hablamos de error de medición diferencial, como en la figura 7, donde los errores de medición de la exposición y el desenlace (O_E y O_D) son independientes, ya que todos los caminos entre ambas variables están bloqueados por factores de colisión. Por el contrario, en el ejemplo de la figura 8, el error de medición es no diferencial, ya que ni el error de medición de la exposición depende del desenlace ni el error de medición del desenlace depende de la exposición; pero ambos errores de medición son dependientes, ya que comparten alguna causa (C_{ED}). Esto sucede, por ejemplo, en estudios de desenlaces ocurridos en el pasado donde tanto la medición de la exposición como la del desenlace, realizadas mediante entrevistas a los participantes, están sujetas al error de memoria.

Figura 8. Sesgo de información no diferencial con errores de medición dependientes. (16)

En general, los errores de medición producen un sesgo de información, salvo cuando E y D no están asociados y los errores de medición son independientes y no diferenciales. Los sesgos de información pueden ir en cualquier dirección, aumentando o reduciendo la fuerza real de la asociación, y su magnitud depende de la magnitud del error de medición.

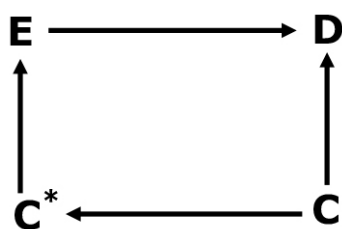
Figura 9. Error de medición en un factor de colisión. (16)

Cuando el error de medición afecta a un factor de colisión, como en la figura 9, ajustar por S^* inducirá por lo general un sesgo de selección en la estimación de la asociación entre la exposición y el desenlace (E y D), de la misma forma que ocurriría ajustando por S, ya que S^* es un descendiente de C y, por tanto, un efecto común (factor de colisión) de E y D.

Figura 10.a. Error de medición en un factor de confusión. (16)

Por último, cuando el error de medición afecta a un factor de confusión, como en la figura 10.a, ajustar por C^* no garantiza de forma general el control completo de la confusión, salvo cuando el efecto de C en la exposición está mediado completamente por C^* , como en la figura 10.b, donde ajustar por C^* es suficiente para controlar la confusión.

Ejemplo: suponga que la presión arterial alta (C) afecta a la probabilidad de recibir tratamiento antihipertensivo (E) y a la de sufrir un accidente cerebrovascular (D). Los médicos y los pacientes, sin embargo, toman sus decisiones sobre el tratamiento basándose en la presión arterial registrada (C^*), la única información sobre la presión arterial a la que tienen acceso, en lugar de la presión arterial real (C). C^* media completamente el efecto de C en E porque cualquier componente de C que no ha sido capturado por C^* permanece desconocido y no puede influir en la decisión de administrar el tratamiento. Por tanto, el camino de puerta de atrás entre E y D puede bloquearse ajustando tanto por C como por C^* . (16) En el resto de los casos, el error de medición también puede ser visto como confusión residual (no medida), si pensamos en C como una variable no medida y en C^* como una sustituta de esa variable.

Figura 10.b. Error de medición en un factor de confusión. (16)

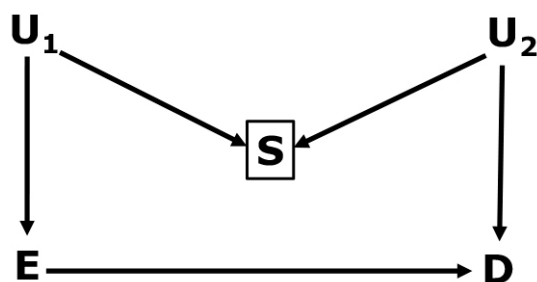
En general, cuando un factor de confusión no se puede medir adecuadamente, es mejor ajustar por una variable sustituta (surrogate) que usar el estimador sin ajustar. Cuanto menor sea el error de medición, más fuerte será la asociación entre C y C*, y mayor el control de la confusión que se conseguirá al ajustar por C*. (3)

4. Control de la confusión

Como hemos mencionado al inicio del capítulo, las aproximaciones tradicionales para controlar la confusión se basaban en las asociaciones estadísticas observadas entre variables en la población de estudio y en la aplicación de los criterios convencionales para considerar que una variable es un factor de confusión (la variable está asociada con la exposición en la población, está asociada con el desenlace en la población no expuesta, y no es un mediador en la relación entre exposición y desenlace). Al analizar la estructura causal del problema, mediante la representación de las relaciones entre variables en un gráfico causal, se pone de manifiesto que estas estrategias pueden resultar inadecuadas. Por ejemplo, en la figura 11 no hay ningún camino por la puerta de atrás abierto entre E y D, porque el único posible está bloqueado en el factor de colisión S, por tanto, no existe confusión. Efectivamente, el gráfico causal muestra la ausencia de causas comunes de E y D. Sin embargo, S cumple los criterios convencionales para ser considerado un factor de confusión, ya que está asociada con la exposición (comparten la causa U_1), está asociada con el desenlace en la población no expuesta (comparten la causa U_2), y no es un mediador en la relación entre exposición y desenlace. Por tanto, siguiendo la aproximación tradicional, al ajustar por S en ausencia de confusión, se genera la confusión que pretendíamos evitar. Por la forma del gráfico causal, la confusión inducida por este ajuste estadístico indebido se denomina sesgo M (M-bias). (17) En este gráfico, hablamos de confusión porque al ajustar por S hemos abierto un camino por la puerta de atrás entre E y D a través de las causas comunes de E y S (U_1) y de S y D (U_2). Hay autores que hablan de sesgo

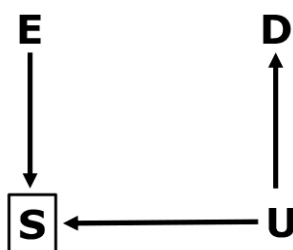
de selección en este caso. Lo importante, al margen de la terminología usada para referirnos a los diferentes sesgos, es la capacidad para identificarlos y controlarlos. A modo de ejemplo, pueden resultar afectados por este tipo de sesgo los estudios de casos y controles hospitalarios diseñados para identificar factores de riesgo de Covid-19 o los estudios para identificar factores pronósticos en estudios de casos y controles con muestras de personas que han dado positivo a un test de Covid-19. (18)

Figura 11. Sesgo M por ajuste indebido de S en ausencia de confusión. (17)



La figura 12 muestra otro ejemplo donde ajustar por una variable que cumple los criterios convencionales para ser considerada un factor de confusión, genera un sesgo previamente inexistente. En este gráfico, al igual que en el caso de la paradoja de la obesidad, hablamos de sesgo de selección porque el camino no dirigido abierto al ajustar por S no es un camino por la puerta de atrás. Por tanto, los criterios meramente estadísticos para seleccionar variables de ajuste, que no tienen en cuenta la estructura causal del problema, son inadecuados para el control de la confusión.

Figura 12. Sesgo por ajuste indebido de S en ausencia de confusión. (17)

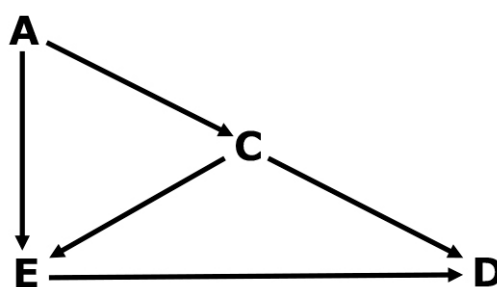


4.1. Conjunto mínimo suficiente de variables

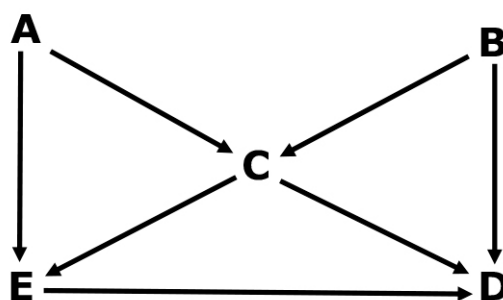
Dado un gráfico causal, asumiendo que la información recogida en el mismo es correcta y completa, podemos identificar las variables por las que es necesario ajustar para controlar la confusión. Un gráfico causal completo debe incluir las variables con un efecto causal en la exposición y el desenlace (factores de confusión), las variables mediadoras del efecto causal y, en su caso, una variable que represente

la selección de los participantes en el estudio. El conjunto de variables suficiente para controlar la confusión, digamos N , será aquel que no contiene ninguna variable descendiente de E y en el que todos los caminos por la puerta de atrás entre E y D están bloqueados al ajustar por N ; es decir, que los grupos de expuestos y no expuestos son intercambiables en cada categoría de N . Este es el denominado criterio de la puerta de atrás (*backdoor criterion*), donde N permite el control de la confusión creada por las causas comunes de E y D , al tiempo que evita inducir una confusión previamente inexistente al ajustar indebidamente por factores de colisión que cumplen los criterios convencionales para ser considerados factores de confusión, como en los ejemplos de las figuras 11 y 12. A partir de las condiciones mencionadas, dado un gráfico causal, podemos aplicar el siguiente test de la puerta de atrás para el control de la confusión (*backdoor test for sufficiency*): Después de ajustar por N , los caminos abiertos entre E y D solo pueden pasar a través de una de las variables de N o de variables que son causas de E o D .

Figura 13.a Conjunto mínimo suficiente de variables para controlar la confusión. (2)



Dado un gráfico causal, decimos que un conjunto de variables N es el mínimo suficiente para controlar la confusión cuando no existe un subconjunto de N suficiente para controlar la confusión. (2) Dado un gráfico causal, pueden existir varios conjuntos de variables suficientes para controlar la confusión. Por ejemplo, en el gráfico causal de la figura 13.a, donde existen dos causas comunes de E y D (A y C), podríamos ajustar por ambas variables, pero como la relación de A con D está mediada enteramente por C , con ajustar únicamente por A o por C sería suficiente. En la figura 13.b, donde existen tres causas comunes de E y D (A , B y C), podríamos ajustar por las tres, pero como las relaciones de A con D y B con E están mediadas enteramente por C , siguiendo el mismo razonamiento, podríamos pensar que con ajustar por C sería suficiente. Sin embargo, al ajustar por C (factor de colisión en este caso), abrimos un camino por la puerta de atrás entre E y D , a través de A y B . Podemos bloquear este camino ajustando por A o por B , por tanto, tendríamos dos posibles conjuntos de variables suficientes para controlar la confusión: A y C o B y C .

Figura 13.b Conjunto mínimo suficiente de variables para controlar la confusión. (2)

Esta forma de proceder es lo que denominamos separación gráfica o separación dirigida (*d-separation*). Decimos que un conjunto de variables N separa otros dos conjuntos E y D o bloquea cualquier camino entre E y D si se cumplen los siguientes criterios:

1. Cada camino abierto entre E y D es interceptado por una variable de N , y
2. Cada camino abierto entre E y D al ajustar por N es interceptado por una variable de N .

Con esta definición, dado un gráfico causal, decimos que N es suficiente para controlar la confusión cuando no contiene ninguna variable descendiente de E o D y separa a E de D en el gráfico obtenido al eliminar todas las flechas que parten de E . Aplicando este criterio al gráfico de la figura 13.b, comprobamos que tanto A y C como B y C separan a E de D ; mientras que C por sí sola no basta para separar a E de D . Esta figura sirve también para ilustrar que un confusor (o un colisionador) en el paso de un camino no es necesariamente un confusor (o un colisionador) en el paso de otro camino. C es un colisionador en $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$ y un confusor en $E \leftarrow C \rightarrow D$.

4.2. Ventajas y limitaciones de los gráficos causales

En los apartados previos, hemos descrito las utilidades de los gráficos causales y las ventajas de su uso frente a otras aproximaciones, pero los gráficos causales también adolecen de ciertas limitaciones, que describimos, junto con sus ventajas, en la tabla 2.

Tabla 2. Algunas ventajas y limitaciones de los gráficos causales

Ventajas	Limitaciones
Para especificar un gráfico causal para un problema en particular, se necesita tener cierto conocimiento de la estructura causal subyacente (contexto). Así, los gráficos causales pueden mostrar la secuencia de eventos causales.	Las asunciones, en particular la de direccionalidad (la causa precede al efecto), generalmente son obtenidas de información ajena a los datos del estudio.
- Llevar a una presentación explícita de algunos de los supuestos (asunciones) que sostiene el investigador. - Permiten visualizar y razonar sobre confusión y sesgo de selección. La confusión se define como asociación de E y D que surge de paso no causal (sesgado) de E a D que comienza con una flecha hacia E y finaliza con una flecha hacia D (camino abierto por la puerta de atrás). - Hay amplio consenso en que los métodos gráficos, como test operativos, han contribuido a clarificar el problema de la confusión.	Las inferencias causales obtenidas siempre dependen de los supuestos particulares codificados en el gráfico causal.
Elaboran la estructura de sesgos (de confusión, selección e información) y hacen éstos explícitos y más manejables.	La práctica totalidad de los estudios epidemiológicos presentan errores de error de medición y confusión no medida.
- Análisis de datos: permiten identificar las variables por las que hay que ajustar. - Proporcionan una justificación explícita para la construcción de conjuntos de variables de ajuste.	- Las flechas NO informan sobre: - La magnitud y sentido (riesgo o protección) de la asociación. - La relación dosis-respuesta. - Incapacidad de manejar la modificación de efecto (no tenemos flechas que vayan a otras flechas).
Contribuyen a desarrollos metodológicos con gráficos causales abstractos.	- Las estructuras en los gráficos causales pueden ser complejas. - Puede haber desacuerdo sobre la estructura causal y se pueden considerar varios gráficos causales diferentes para representar el mismo problema.

Fuente: Elaboración propia.

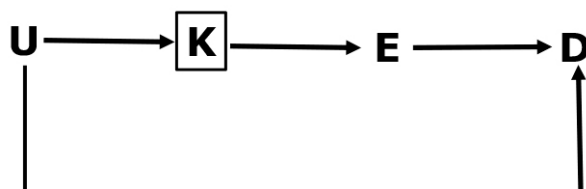
4.3. Principios para el control de la confusión

Como hemos señalado a lo largo de este capítulo, el conocimiento sustantivo, cuando no completo, de las relaciones entre las variables y la estructura causal subyacente al desenlace de interés es un requisito previo para identificar las variables por las que es necesario ajustar para controlar la confusión. El análisis por separado

de cada variable resulta insuficiente e inadecuado para seleccionar las variables por las que es necesario ajustar y los análisis estadísticos por sí solos no permiten determinar el orden temporal de las relaciones ni distinguir entre factores de confusión y mediadores, como en las siguientes situaciones:

- Primero, el ajuste por una variable de confusión podría introducir confusión por otras variables que no son factores de confusión por sí mismas, pero que están relacionadas con la variable por la que hemos ajustado.
- Segundo, el ajuste por una variable que no es un factor de confusión (K) podría eliminar la confusión por una causa no medida del desenlace con la que estuviera relacionada (U), como en el gráfico de la **figura 14**. (19)

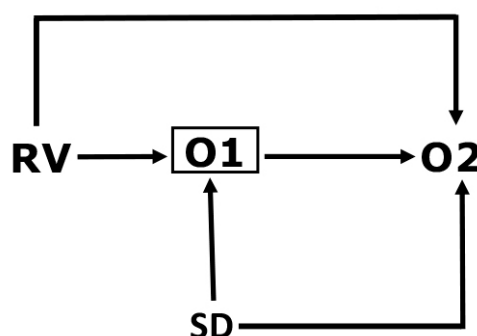
Figura 14. Control de la confusión por ajuste del descendiente de un confusor desconocido. (18)



- Tercero, el ajuste por una variable que, cumpliendo los criterios convencionales para ser considerado factor de confusión, resulta ser un factor de colisión puede inducir una confusión previamente inexistente o un sesgo de selección. Incluso podría explicar completamente una asociación espuria observada.

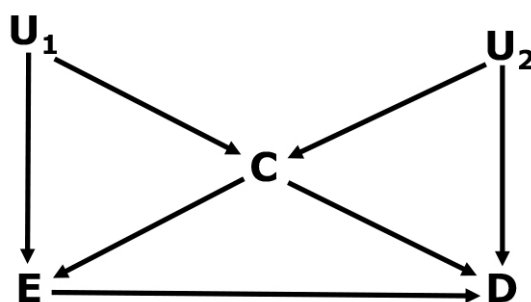
Ejemplo: un reciente estudio muestra que la detección en algunos estudios observacionales de un mayor riesgo de reinfección por la variante Omicron del SARS-CoV-2 en sujetos que habían recibido dosis de refuerzo de la vacuna de la Covid-19 y habían tenido una primera infección post-refuerzo no debe interpretarse como evidencia de un efecto adverso del refuerzo vacunal. Más bien se debe a un sesgo de selección por restringir (condicionar) los análisis a sujetos con una infección previa por Omicron post-vacunación, que es un factor de colisión (efecto común del refuerzo vacunal y de una mayor susceptibilidad a la infección). El mayor riesgo de reinfección en los que recibieron el refuerzo vacunal se debería, por tanto, a que se trataba de sujetos más susceptibles a la infección (**figura 15**). (20)

Figura 15. Sesgo de colisión por condicionar en sujetos con infección previa por la variante Omicron de SARS-Cov 2 (RV: refuerzo vacunal; O1: primera infección; O2: reinfección. SD: susceptibilidad a infección). (19)



- Por último, también puede darse el caso de que una variable sea, a un tiempo, un factor de confusión y un factor de colisión, como en el gráfico de la **figura 16**. En estos casos, hay que tener presente que la magnitud del sesgo introducido por no ajustar cuando existe confusión o ajustar cuando no existe depende de la fuerza de las asociaciones estadísticas entre el potencial factor de confusión y las variables de exposición y desenlace. Además, los estudios de simulación muestran que la magnitud del sesgo M por ajuste de un factor de colisión es pequeña, salvo cuando las asociaciones entre el factor de colisión y las variables no medidas (U_1 y U_2) son muy fuertes, con riesgos relativos por encima de 8. (21) En general, para los escenarios considerados más realistas en diversas simulaciones, el sesgo de confusión fue siempre de mayor magnitud que el sesgo de colisión, por lo que, cuando un posible factor de colisión es en sí mismo un factor de confusión importante, es aconsejable ajustar por esa variable, aun a costa de inducir un sesgo de colisión, ya que este será previsiblemente de menor magnitud que el sesgo de confusión evitado.

Figura 16. Control de la confusión en presencia de un confusor que es, a un tiempo, factor de colisión de dos variables no medidas. (20)



En la práctica, es raro que tengamos un conocimiento completo de las relaciones entre las variables y la estructura causal subyacente al desenlace de interés; pero siempre es preferible tomar las decisiones teniendo en consideración las suposiciones o creencias causales derivadas del conocimiento previo del problema en cuestión, y no meramente basarnos en las asociaciones estadísticas detectadas en los datos a los que tiene acceso el investigador. Ante la necesidad de tomar decisiones con un conocimiento más o menos imperfecto, a veces muy limitado, de la estructura causal del problema, y en vista de que las estrategias utilizadas con frecuencia resultan insuficientes e inadecuadas, VanderWeele ha propuesto unos principios para seleccionar las variables que hay que incluir en un modelo para controlar la confusión, que pretenden evitar los problemas de las alternativas tradicionales. (19) Estos principios, que conforman el denominado criterio causal disyuntivo modificado (*modified disjunctive cause criterion*) son los siguientes:

1. Seleccionar aquellas variables que son causas de la exposición o del desenlace o de ambos.
2. Descartar cualquier variable instrumental.
3. Añadir variables que no satisfacen el principio 1, pero que son buenas sustitutas de causas comunes de la exposición y el desenlace que no han sido medidas.

El propio VanderWeele reconoce que estos principios no son definitivos, pues se trata de una cuestión compleja y abierta a la discusión, con mucha investigación en curso que permitirá afinar las estrategias para un mejor control de la confusión en los estudios epidemiológicos futuros. Además, el criterio causal disyuntivo modificado adolece en cierta medida de la misma limitación que el criterio de la puerta de atrás, ya que para aplicarlo necesitamos información sobre las relaciones entre variables que no siempre está a nuestro alcance por falta de conocimientos, especialmente los relativos a los criterios que debe cumplir una variable instrumental, lo que impide tener certeza acerca de si la variable en cuestión es o no instrumental. Mientras no se resuelvan estas dificultades, resulta muy recomendable la realización de múltiples análisis de sensibilidad, ajustando por diferentes conjuntos de variables, ya que sus resultados, de ser compatibles, nos proporcionarán mayor confianza en la verosimilitud de los mismos.

5. Referencias

1. Szklo M & Nieto FJ. Epidemiology: beyond the basics. Fourth edition. Burlington: Jones & Bartlett Learning. 2019.
2. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37-48.
3. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155:176-84.
4. Lipsky AM, Greenland S. Causal Directed Acyclic Graphs. *JAMA*. 2022;327:1083-4.
5. Hernández-Díaz S, Schisterman E, Hernán MA. The weight birth “paradox” uncovered. *Am J Epidemiol*. 2006;164:1115-20.
6. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *British Medical Journal*. 1950;2:739-48.
7. Hammond EC, Horn D. The relationship between human smoking habits and death rates: A follow-up study of 187,766 men. *Journal of American Medical Association*. 1954;155:1316-28.
8. Hecht SS, Szabo E. Fifty years of tobacco carcinogenesis research: from mechanisms to early detection and prevention of lung cancer. *Cancer Prev Res (Phila)*. 2014;7:1-8.
9. Wild CP, Weiderpass E, Stewart BW, editors (2020). World Cancer Report: Cancer Research for Cancer Prevention. Lyon, France: International Agency for Research on Cancer. Available from: <http://publications.iarc.fr/586>
10. Hainer V, Aldhoon-Hainerová I. Obesity paradox does exist. *Diabetes Care*. 2013;36(Suppl 2):S276-81.
11. Wolfram Doehner, Erland Erdmann, Richard Cairns, Andrew L Clark, John A Dormandy, Ele Ferrannini, Stefan D Anker. Inverse relation of body weight and weight change with mortality and morbidity in patients with type 2 diabetes and cardiovascular co-morbidity: an analysis of the PROactive study population. *Int J Cardiol*. 2012;162:20-6.
12. Lajous M, Bijon A, Fagherazzi G, Boutron-Ruault MG, Balkau B, Clavel-Chapelon F, Hernán MA. Body mass index, diabetes, and mortality in French women: explaining away a “paradox”. *Epidemiology*. 2014;25:10-4.
13. Holmberg MJ, Andersen LW. Collider Bias. *JAMA*. 2022;327:1282-3.

14. Greenland S. Response and follow-up bias in cohort studies. *Am J Epidemiol.* 1977;106:184-7.
15. Hernán MÁ. Invited Commentary: Selection Bias Without Colliders. *Am J Epidemiol.* 2017;185:1048-50.
16. Hernán MA, Robins JM (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
17. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology.* 2003;14:300-6.
18. Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, *et al.* Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications.* 2020;11:5749.
19. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol.* 2019;34:211-9.
20. Monge S, Pastor-Barriuso R, Hernán MA. The imprinting effect of covid-19 vaccines: an expected selection bias in observational studies. *BMJ.* 2023;381:e074404.
21. Liu W, Brookhart MA, Schneeweiss S, Mi X, Setoguchi S. Implications of M bias in epidemiologic studies: a simulation study. *Am J Epidemiol.* 2012;176:938-48.

Capítulo 8.

Encuestas de salud poblacionales

IÑAKI GALÁN LABACA

1. Introducción
2. Modelos de encuestas de salud poblacionales
3. Muestreo
4. Métodos de entrevista
5. Contenidos del cuestionario
6. Tipos de errores en las encuestas
7. Tasa de respuesta
8. Análisis de datos con diseños de muestras complejas
9. ¿Se pueden estudiar asociaciones?
10. Encuestas de salud repetidas
11. Encuestas panel
12. Referencias

1. Introducción

Las encuestas de salud proporcionan información sobre múltiples dimensiones que son claves para la planificación sanitaria y la toma de decisiones en salud pública, abarcando desde la prevención y control de los principales problemas de salud hasta la evaluación de intervenciones sanitarias. En este capítulo nos centraremos en las encuestas de salud poblacionales y más concretamente en las encuestas generales de salud, donde la información que se recoge es muy amplia y atiende a múltiples objetivos, por lo que se denominan también encuestas multipropósito.

Las encuestas de salud han estado en continua evolución desde sus inicios, al final del siglo XIX, con el aporte metodológico de numerosas disciplinas como la demografía, sociología, economía, psicología, estadística, y la medicina y salud pública (1). A mediados del siglo XX, la menor importancia de las enfermedades infecciosas para la salud de la población, y el incremento en la esperanza de vida como consecuencia del descenso de la mortalidad, conduce a prestar un mayor interés en la morbilidad. Se pone mucho énfasis en esa época en la identificación de la prevalencia y las causas de las enfermedades crónicas, discapacidad, comportamientos de salud, y en la utilización de la atención sanitaria.

Actualmente las encuestas de salud poblacionales forman parte de una compleja red de sistemas de información, aportando información sobre el estado de salud, factores de riesgo, seguimiento de las recomendaciones preventivas, y uso de los dispositivos asistenciales del sistema sanitario.

Básicamente son estudios transversales, es decir, realizan una fotografía en un periodo temporal del estado de salud de una población. Como generalmente son estudios repetidos en el tiempo, sirven para conocer tendencias y estimar proyecciones, así como para evaluar intervenciones. También identifican grupos que requieren cuidados especiales, y sirven para planificar la provisión de servicios. Además, se usan para establecer relaciones o asociaciones entre variables, así como generar hipótesis. Sin embargo, para muchos temas concretos las encuestas pueden diseñarse como paneles, donde el objetivo es determinar cambios en el tiempo, siendo estudios longitudinales con una duración determinada (2). Además, cada vez es más frecuente que sus datos se puedan cruzar con la mortalidad, u otros indicadores de salud, por lo que el original diseño transversal se convierte en una cohorte con base poblacional (3,5).

Las encuestas de salud pueden fundamentarse en información autodeclarada por el entrevistado, o incluir medidas objetivas de determinados parámetros de salud. En ese caso, se consideran exámenes de salud, que pueden realizarse de forma combinada con la encuesta de salud, utilizando la misma muestra o en una submuestra, o realizarse de forma independiente.

2. Modelos de encuestas de salud poblacionales

La mayor experiencia en el uso de encuestas de salud proviene de Estados Unidos. En 1935-36 se realiza la primera Encuesta Nacional de Salud de Estados Unidos (National Health Interview Survey [NHIS]) (6), no diseñada en aquel momento como una fuente de información continua de salud. Desde 1957 se conduce de forma continua. El diseño actual, implementado en 1997, tiene preguntas básicas o centrales y preguntas complementarias. Las preguntas centrales permanecen prácticamente sin cambios y permiten el análisis de tendencias y la combinación de datos de más de un año para aumentar la muestra e incrementar el poder estadístico. El módulo básico del hogar recoge de todos los miembros información sociodemográfica y sobre el estado de salud, limitaciones funcionales, lesiones, acceso y utilización de atención médica, seguro médico. En cada hogar, se selecciona un adulto, y un niño, que tienen cuestionarios independientes, centrados en el estado de salud, comportamientos o estilo de vida y uso del sistema sanitario. Además, tiene suplementos con preguntas adicionales que no se recogen todos los años pero que se repiten con periodicidad regular para monitorizarlas.

Complementariamente, la National Health and Nutrition Examination Survey (NHANES), evalúa el estado de salud y nutricional de la población infantil y adulta de Estados Unidos (7). Comenzó en 1960 y desde 1999 se convirtió en una encuesta continua realizando el examen de salud en una muestra de 5000 personas cada año. Numerosos aspectos de salud son recogidos a través de entrevistas personales, exámenes físicos y test de laboratorios.

La Encuesta de Salud de Inglaterra (Health Survey for England [HSE]) (8) se puso en marcha en el año 1991 con una muestra de menores de 16 años (4000) y otra dirigida a población adulta (16 000), realizándose anualmente a partir de ese año. Además de información de las características del hogar y diversos aspectos de salud y uso de servicios sanitarios de los individuos, incluye un examen de salud que recoge medidas de presión arterial, peso y talla, y una muestra de sangre.

En España, la Encuesta Nacional de Salud se realiza por primera vez en 1987, para repetirse de forma discontinua en 1993, 1995, 1997, y 2001. Desde 2003 se realiza cada cuatro años mediante una colaboración entre el Ministerio de Sanidad y el Instituto Nacional de Estadística (INE) (9). Se basa en una entrevista personal sin examen de salud, por lo que toda la información es autodeclarada. De forma complementaria, la European Health Interview Survey (EHIS), coordinada por Eurostat y regulada por Reglamento (EC) 1338/2008 y Reglamento de la Comisión 141/2013 (10), se realiza también cada cuatro años no coincidentes con la Encuesta Nacional de Salud. El cuestionario para España fue adaptado para permitir la comparación con los principales indicadores de la Encuesta Nacional de Salud. Por lo tanto, cada dos años se recogen datos de una de las dos encuestas.

3. Muestreo

El objetivo principal de las encuestas de salud, recoger información sobre múltiples aspectos que son claves para la planificación sanitaria, maximiza la importancia de que los datos sean representativos del total de la población.

Conviene distinguir los siguientes conceptos que a veces se hacen un poco confusos:

- Población de estudio: es el conjunto de elementos sobre los que se investiga. En el caso de que, por ejemplo, no se incluya a la población institucionalizada (residencias, prisiones, etc.), los datos no se podrán inferir a estas poblaciones excluidas del estudio. Normalmente en las encuestas poblacionales se suelen excluir.
- Marco de muestreo: es el listado de donde se seleccionará la muestra.
- Muestra: son todas las unidades incluidas en el estudio y que serán objeto del análisis.

3.1. Tipos de muestreo

Inicialmente se diferencian en muestreos probabilísticos y no probabilísticos, según las unidades muestreadas tengan o no una probabilidad conocida distinta de cero de ser seleccionadas en el estudio. En el primero de ellos el objetivo es inferir los datos a la población por lo que la representatividad es un aspecto clave, mientras que en el segundo no se busca la representatividad (11).

El muestreo probabilístico incluye los siguientes tipos (12):

- Muestreo aleatorio simple: todas las unidades del marco de muestreo tienen una misma probabilidad de ser seleccionadas en la muestra. Se asigna un número a cada unidad y mediante un procedimiento de generación de números aleatorios se seleccionan las unidades.

Como gran ventaja, requiere poco conocimiento de la distribución de la población. Como desventajas, no es muy eficiente y puede no capturar bien a ciertos grupos de interés.

- Muestreo sistemático: es similar al aleatorio simple pero utiliza una aproximación sistemática para realizar la selección. El marco de muestreo se numera secuencialmente, seleccionando un número aleatorio para el arranque, calculando el intervalo de salto en función del tamaño muestral estimado.

Requiere poder identificar a las unidades para ordenar a la muestra. Este orden puede crear sesgos al realizar la selección. Al igual que en el método anterior, no es muy eficiente y puede no capturar bien a ciertos grupos de interés.

- Muestreo estratificado: el marco de muestreo se divide en grupos en función de los objetivos del estudio. En las encuestas de salud poblacionales la muestra suele estratificarse por unidades geográficas, pero por ejemplo en las de salud escolar por tipos de colegio (públicos o privados), o en las de uso de servicios sanitarios por tipos de servicios (tipo de hospital, por ejemplo). La selección puede ser proporcional o no, esto quiere decir que algunos grupos de unidades tienen mayor probabilidad de salir elegidas respecto a un muestreo aleatorio simple mientras que otras tienen menor probabilidad. Por ejemplo, en la Encuesta Nacional de Salud y Encuesta Europea de Salud, las comunidades autónomas más pequeñas tienen más tamaño muestral que la que les correspondería según la distribución de su población. Se utiliza para poder inferir los datos al conjunto de este grupo o estrato con un mínimo de poder estadístico.

Como ventajas, aumenta la precisión de las estimaciones y posibilita el análisis de ciertos grupos de interés. Como desventajas, requiere conocer la distribución de la población.

- Muestreo de conglomerados: las unidades muestrales están habitualmente agrupadas de forma natural en unidades más grandes. Por ejemplo, los escolares en los colegios, las personas que demandan atención sanitaria en centros de salud o en hospitales, o a nivel poblacional el área de residencia como las secciones censales, barrios, municipios, etc.

Como ventajas, destacar la disminución de costes del trabajo de campo (las unidades en lugar de estar dispersas, están concentradas o agrupadas). Además, se puede incorporar al análisis de los datos ciertas características de las unidades de agrupación que actuarán como variables explicativas. Como desventaja, disminuye la precisión de las estimaciones y el análisis de los datos es más complejo.

En la práctica, las encuestas de salud poblacionales utilizan una combinación de los métodos anteriores, principalmente un muestreo estratificado y de conglomerados para seleccionar los hogares, y una selección aleatoria simple dentro del hogar para seleccionar a los individuos. Estos procedimientos se desarrollan en varias etapas, por lo que se denominan muestreos polietápicos. Un ejemplo de ello es el muestreo seguido en la ENSE y que describiremos a continuación.

El muestreo no probabilístico también tiene varios métodos:

- Muestreo por cuotas, donde los participantes son seleccionados según unas características que tienen la misma distribución que las de la población general.
- Muestreo de conveniencia u oportunistico, donde la muestra se selecciona de acuerdo a criterios de facilidad en el reclutamiento.
- Muestreo en bola de nieve, donde las personas entrevistadas van invitando a otras para participar en el estudio. Generalmente se usa en poblaciones de difícil acceso.
- Muestreo intencional, seleccionando los investigadores poblaciones específicas que aportan información que no podría ser obtenida de otras fuentes (médicos, enfermeras, personas encargadas de la formulación de las políticas, etc.).

3.2. Muestreo de la Encuesta Nacional de Salud

Como se ha comentado anteriormente, las encuestas de salud poblacionales utilizan para la selección de las unidades un muestreo complejo dividido en varias etapas, y, por lo general, es estratificado y de conglomerados. Vamos a ver el ejemplo de la ENSE de 2016-17, donde el INE utiliza el siguiente procedimiento (13).

Es un muestreo dividido en tres etapas, estratificado y de conglomerados. Las unidades de primera etapa son las secciones censales (conglomerados). Las unidades de segunda etapa son las viviendas familiares principales. Dentro de cada hogar se selecciona a un adulto (15 o más años) para cumplimentar el cuestionario de adultos y en caso de que hubiera menores (de 0 a 14 años) se selecciona, asimismo, un menor para cumplimentar el cuestionario de menores. El marco utilizado para

la selección de la muestra de unidades de primera etapa es un marco de áreas formado por la relación de secciones censales existentes. Para las unidades de segunda etapa se utiliza la relación de viviendas familiares principales en cada una de las secciones seleccionadas para la muestra. Las unidades de tercera etapa se seleccionan a partir de la relación de personas encuestables en la vivienda en el momento de la realización de la entrevista. Las unidades de primera etapa se agrupan en estratos de acuerdo con el tamaño del municipio al que pertenecen las secciones. Se consideran los siguientes estratos: Estrato 1: municipios de más de 500 000 habitantes. Estrato 2: Municipio capital de provincia (excepto los anteriores). Estrato 3: municipios con más de 100 000 habitantes (excepto los anteriores). Estrato 4: municipios de 50 000 a 100 000 habitantes (excepto los anteriores). Estrato 5: municipios de 20 000 a 50 000 habitantes (excepto los anteriores). Estrato 6: municipios de 10 000 a 20 000 habitantes. Estrato 7: municipios con menos de 10 000 habitantes. La muestra se distribuye entre comunidades autónomas, asignando una parte uniformemente, y, otra, de forma proporcional al tamaño de la comunidad. Para cada comunidad autónoma se diseña una muestra independiente que la representa, por ser uno de los objetivos de la encuesta facilitar datos con este nivel de desagregación.

3.3. Criterios para estimar el tamaño muestral

El cálculo del tamaño muestral se guía por numerosos criterios (1):

- Dado que las encuestas de salud poblacionales miden muchos indicadores, hay que identificar las variables de estudio principales para determinar el nivel de confianza deseado decidiendo un rango de error tolerable.
- La necesidad de realizar estimaciones para subgrupos poblacionales, valorando el nivel de desagregación geográfica y de estratos sociodemográficos (edad, sexo, nivel de estudios, etc.) que queremos realizar. Por ejemplo, en la ENSE los datos son representativos de la comunidad autónoma, aunque la precisión de las estimaciones para muchas de ellas, en especial cuando estratificamos por variables sociodemográficas, puede resultar pequeña.
- El tamaño muestral efectivo basado en el efecto de diseño, es decir, en qué grado el error estándar estimado a partir de una muestra compleja (especialmente en el muestreo de conglomerados) difiere del estimado mediante un aleatorio simple. Por ejemplo, si el efecto de diseño es de 1,6, quiere decir que el error estándar va a ser 1,6 veces más grande, por lo que tendremos que multiplicar este efecto de diseño al tamaño muestral estimado mediante un procedimiento aleatorio simple. Además, hay que considerar la tasa de respuesta. Si se calcula que puede ser del 70 %, hay que dividir

además la cifra anterior por 0,70. Por ejemplo, si la muestra original fuera de 15 000 individuos basado en un cálculo para muestreo aleatorio simple: 1) Teniendo en cuenta el efecto de diseño de 1,6: $15\,000 \times 1,6 = 24\,000$; 2) Teniendo en cuenta la tasa de respuesta: $24\,000 / 0,70 = 34\,285$, que sería el tamaño muestral acorde con los criterios del estudio.

- El presupuesto disponible para realizar la encuesta.

4. Métodos de entrevista

4.1. Encuesta personal

La encuesta personal o cara a cara es la técnica más frecuentemente utilizada en encuestas de salud poblacionales. Se realiza en presencia de un entrevistador bien a través de una encuesta en papel o mediante encuesta asistida por ordenador (CAPI —Computer Assisted Personal Interviewing—). La entrevista mediante CAPI permite ahorrar tiempo y facilita un mejor registro de la información. La Encuesta Nacional de Salud se realiza mediante esta técnica desde 2011-12 y la Encuesta Europea de Salud para España desde 2009-10.

Ventajas	Inconvenientes
• Elevada respuesta • Facilidad de cooperación y resolución de dudas • Evita influencia de otras personas • Posibilita realizar entrevistas largas • Pueden mostrarse materiales, fichas, etc. • Pueden obtenerse datos secundarios del ambiente de la entrevista	• Caras • Más lentas • El entrevistador puede influir en las respuestas • Dificil acceso a poblaciones ocupadas • Mayor riesgo de sesgo de conveniencia social

4.2. Encuesta telefónica

La encuesta telefónica, generalmente asistida por ordenador (CATI —Computer Assisted Telephone Interviewing—), utiliza un sistema informático que selecciona y realiza aleatoriamente las llamadas y permite al entrevistador complementar el cuestionario con ayuda de un ordenador. El entrevistador lee el cuestionario desde la pantalla e introduce la información directamente en el ordena-

dor. Poco frecuente en encuestas de salud es el uso de la técnica ACTI (Automated Computer Telephone Interviewing), que utiliza una voz interactiva que reemplaza al entrevistador y se registra la información mediante palabras y frases reconocibles por el ordenador.

Ventajas	Inconvenientes
• Más barata • Muy rápida • Permite entrevistar a determinadas poblaciones con dificultades de acceso • Mejor disponibilidad a responder ciertas preguntas • Menos entrevistadores • Buen control del trabajo de campo	• Depende de la cobertura telefónica • No pueden exhibirse fichas, materiales, etc. • Tiempo de entrevista más corto • No apropiado para tratar temas complejos que requieran de explicaciones o guías • Mayor riesgo de sesgo de conveniencia social

Esta técnica de entrevista telefónica se utiliza en la encuesta “Behavioural Risk Factor Surveillance System (BRFSS)” en Estados Unidos, que monitoriza los principales factores de riesgo de las enfermedades crónicas (14). En España es utilizada desde 1995 en el Sistema de Vigilancia de Factores de Riesgo de Enfermedades Crónicas (SIVFRENT) de la Comunidad de Madrid (15).

Algunas encuestas usan directorios como marco de muestreo. Es el caso de las bases de datos de telefonía fija, muy utilizados en nuestro medio. La gran limitación es que la cobertura de teléfono fijo en España era en 2019 del 75 % a nivel nacional con grandes variaciones entre comunidades autónomas, oscilando entre el 52,5 % de Murcia y el 87,4 % en Madrid (16). Además, los más jóvenes, emigrantes, personas de baja posición socioeconómica, tienen menos cobertura de telefonía fija, lo que puede incurrir en un sesgo de selección. Sin embargo, cuando la cobertura es alta, los resultados de los principales indicadores de salud respecto a los obtenidos por entrevista cara a cara son muy similares (17).

4.3. Encuesta postal

El entrevistado cumplimenta el cuestionario directamente y no existe comunicación directa entre el encuestado y el entrevistador. El número de preguntas que se formulan debe ser breve, y las preguntas tienen que ser “cerradas” y sencillas de cumplimentar.

Ventajas	Inconvenientes
• Muy barata • Gran accesibilidad • Flexibilidad de tiempo para realizarla • Favorece el anonimato • Minimiza el sesgo de conveniencia social	• Baja tasa de respuesta • El cuestionario tiene que ser muy breve • Otras personas pueden influir en la respuesta • Exclusión de personas con dificultad para leer las preguntas

4.4. Encuesta online

Este tipo de encuesta consiste en “colocar” un cuestionario en una página web o en enviarlo a los correos electrónicos de un panel predefinido. Tiene las mismas características que la encuesta postal, excepto el medio en que se desarrolla (Internet).

Ventajas	Inconvenientes
• Muy barata • Muy rápida • Permite utilizar medios audiovisuales durante la encuesta • Flexibilidad de tiempo para realizarla • Favorece el anonimato • Minimiza el sesgo de conveniencia social	• Depende de la cobertura de Internet • Baja tasa de respuesta • Se puede mentir más fácilmente • El cuestionario tiene que ser muy breve • Otras personas pueden influir en la respuesta • Exclusión de personas con dificultad para leer las preguntas • Dificil control de que la persona que responde es la adecuada o realiza varias entrevistas

Un ejemplo de esta técnica en España es el Health Behaviour School-aged Children (HBSC), un estudio sobre las conductas saludables de los jóvenes escolarizados entre los 11 y los 18 años, auspiciado por la Organización Mundial de la Salud (OMS) en el que participan casi 50 países occidentales (18).

También hay que destacar las técnicas mixtas, por ejemplo, las encuestas con entrevista personal cara a cara o postal con seguimiento telefónico, cuyo uso está cada vez más extendido (19).

4.5. Informantes “proxy”

En algunas situaciones no siempre es posible entrevistar a la persona seleccionada directamente. Por ejemplo, en colectivos especiales como niños, personas muy mayores o con discapacidad, o en general cuando los individuos tienen limitaciones de capacidad cognitiva o física, se necesita una asistencia parcial o completa para contestar el cuestionario. Estos informantes “proxies” suelen ser los

padres en el caso de cuestionarios dirigidos a la población infantil, y las esposas, maridos o convivientes que puedan tener buena información de la persona seleccionada (1).

Se pueden clasificar en dos categorías:

1. El “proxy” por necesidad, que se refiere a la situación en la cual la persona seleccionada no puede realizar la entrevista por problemas de capacidad física o mental; y
2. El “proxy” de conveniencia, que sustituye al entrevistado original no por un problema de incapacidad de realizar la entrevista por problemas físicos o mentales sino por la dificultad de contactar de forma efectiva para realizar la entrevista. La sustitución por un informante “proxy” reduce costes al evitar repetir contactos para intentar conseguir entrevistar a la persona seleccionada originalmente. No obstante, esta segunda opción debería limitarse todo lo posible (1).

Los investigadores deben valorar el uso de los informantes “proxy”, teniendo en cuenta el equilibrio entre la reducción del sesgo de selección que se puede incurrir al sustituir a las personas con problemas para realizar la entrevista, y la calidad de la información debido a la dificultad de registrar correctamente ésta debido a problemas de inexactitud en las respuestas, especialmente en aquellas preguntas relacionadas con el estado de salud y limitaciones funcionales.

Por ejemplo, en la ENSE 2016-17, el 2,7 % de las entrevistas fueron cumplimentadas por informantes “proxy”, el 3,3 % debido a que la persona seleccionada se encontraba ingresado en un hospital, un 76,9 % por discapacidad o enfermedad grave, y un 19,8 % a causa de dificultades con el idioma. Los hijos, los cónyuges o parejas, y los padres, fueron, por este orden, los informantes según la relación familiar con la persona originalmente seleccionada.

5. Contenidos del cuestionario

5.1. Ejemplo de contenidos de la Encuesta Nacional de Salud

Como hemos comentado previamente, las encuestas de salud poblacionales son encuestas multipropósito. Por ejemplo, los contenidos del cuestionario de la Encuesta Nacional de Salud 2016-17 dirigida a población adulta son los siguientes:

- **Módulo de variables socioeconómicas.**

■ Módulo de estado de salud:

- ◆ Estado de salud general. Percepción que tiene la persona sobre su estado de salud general en los últimos doce meses.
- ◆ Enfermedades crónicas o de larga duración.
- ◆ Limitaciones por problemas de salud para realizar las actividades habituales. Grado de limitación y tipo de problema.
- ◆ Enfermedades y problemas de salud. Se trata de investigar qué tipo de enfermedades o problemas de salud de larga duración ha padecido la población alguna vez, cuáles ha padecido en los últimos doce meses y si han sido diagnosticados por un/a médico/a.
- ◆ Accidentalidad.
- ◆ Restricción de la actividad.
- ◆ Limitaciones físicas, sensoriales y cognitivas.
- ◆ Limitaciones para la realización de las actividades básicas de la vida diaria.
- ◆ Limitaciones para la realización de actividades instrumentales de la vida diaria, relacionadas con el hogar.
- ◆ Dolor.
- ◆ Salud mental.
- ◆ Condiciones de trabajo.

■ Módulo de asistencia sanitaria.**■ Última vez que consultó al médico general o médico de familia.**

- ◆ Número de veces que ha consultado a médico general o médico de familia en las últimas cuatro semanas.
- ◆ Última vez que consultó a un especialista.
- ◆ Número de veces que ha consultado a un especialista en las últimas cuatro semanas.
- ◆ Lugar de la última consulta realizada en las cuatro últimas semanas.
- ◆ Motivo de la consulta.
- ◆ Tiempo que pasó desde que empezó a notarse enfermo/a o sintió que tenía un problema de salud hasta que pidió cita para consulta y tiempo que pasó desde que pidió cita hasta que fue atendido/a.
- ◆ Realización de análisis en los últimos 12 meses.
- ◆ Realización de pruebas diagnósticas en los últimos 12 meses.
- ◆ Visitas en los últimos 12 meses a un Homeópata, Acupuntor, Naturista u Otros profesionales de la medicina alternativa.

- ♦ Consultas a estomatología, dentista e higienista dental.
- ♦ Hospitalizaciones, servicios de urgencia.
- ♦ Seguro sanitario.
- ♦ Consumo de medicamentos.
- ♦ Tipo de medicamento consumido en las últimas dos semanas y si fue recetado por un médico.
- ♦ Prácticas preventivas generales.
- **Módulo de determinantes de la salud.**
 - ♦ Características físicas: peso y talla.
 - ♦ Actividad física.
 - ♦ Frecuencia de consumo de 15 alimentos.
 - ♦ Higiene dental.
 - ♦ Consumo de tabaco.
 - ♦ Exposición al humo de tabaco.
 - ♦ Consumo de bebidas con alcohol.
 - ♦ Apoyo social.
 - ♦ Cuidado a otras personas con problemas de salud.

5.2. Validez de los contenidos

Lógicamente los contenidos atienden a los objetivos de la encuesta. Normalmente se desearía introducir más contenidos; sin embargo, alargar mucho el cuestionario puede ser contraproducente ya que el entrevistado pierde capacidad de concentración y las respuestas pueden ser muy inexactas. Además, hay que evaluar muy detenidamente la validez de las preguntas formuladas.

Como ejemplo, vamos a valorar la validez de contenidos de la ENSE 2016-17 con los siguientes indicadores: 1) estimación del índice de masa corporal a través del peso y talla autodeclarados; 2) medición del estado de salud general; 3) medición de la salud mental.

- El índice de masa corporal (IMC) (kg/m^2) se estima a partir del peso y la talla percibidos o autodeclarados por el entrevistado: “¿Podría decirme cuánto mide, aproximadamente, sin zapatos?”, “¿Y cuánto pesa aproximadamente, sin zapatos ni ropa?”. En las encuestas en las que no se miden de forma objetiva el peso y la talla, la única forma de estimar el IMC es de forma autodeclarada. Sin embargo, la validez de esta forma de medir el IMC es muy limitada, con una evidencia científica muy consistente del

grado de error, que conlleva una importante subestimación de la obesidad y, en menor medida, del sobrepeso (20).

- El estado de salud general se pregunta de la siguiente manera: En los últimos doce meses, ¿diría que su estado de salud ha sido muy bueno, bueno, regular, malo, muy malo? Es un indicador simple y fácil de administrar como evaluación de salud general, por lo que su incorporación a las encuestas de salud está muy extendida, y se considera que es un buen predictor de la mortalidad y de otras medidas de salud objetivas de morbilidad, como las hospitalizaciones y el uso del sistema de salud. Incluso al ajustar por variables objetivas de salud y sociodemográficas, esta asociación no pierde su poder estadístico por lo que se entiende que no solo es una medida que tiene en cuenta el estado de salud biológico sino que incluye aspectos de salud que son relevantes y no están incluidos en otras medidas de salud (21).
- La salud mental se estima mediante el “General Health Questionnaire” de 12 preguntas (GHQ-12). Consiste en 12 ítems que evalúan la gravedad de diversos problemas en las últimas semanas: pérdida de concentración, pérdida de sueño, desempeño de un papel útil en la vida, toma de decisiones, sensación de agobio y tensión, superación de dificultades, disfrute de las actividades cotidianas, afrontamiento de los problemas, sensación de depresión, pérdida de confianza en uno mismo, duda sobre la propia valía y sensación de felicidad. Se puntúa utilizando una escala Likert con 4 categorías de respuesta: “mucho menos que lo habitual” y “lo habitual” (0 puntos), “más que lo habitual” y “mucho más que lo habitual” (1 punto). Se identifican individuos con posible trastorno mental con una puntuación de 3 o más. Ha sido validado en España y es uno de los instrumentos más utilizados para identificar problemas de salud mental en el ámbito de la salud pública, además de test de screening en el ámbito clínico (22).

6. Tipos de errores en las encuestas

La grandeza de trabajar con muestras probabilísticas, es que, a partir de preguntas a unos cientos o miles de personas, podemos hacer estimaciones precisas de la distribución de importantes indicadores de salud en poblaciones de millones de personas. Sin embargo, estas estimaciones están sujetas a errores de muestreo, además de otras fuentes de error que hay que conocer con el máximo detalle.

Las fuentes de error más importantes son:

- **Error de cobertura:** ocurre cuando no hay una correspondencia perfecta entre la población objeto de estudio y el marco de muestreo del que se extrae una muestra. Esto significa que algunos miembros de la población no tienen probabilidad de salir elegidos.
- **Error de no respuesta:** los participantes son distintos a los seleccionados inicialmente.
- **Error de medición:** preguntas que no recogen la información objeto de estudio, respuestas inexactas muchas veces resultado de preguntas mal formuladas, efecto de los entrevistadores.
- **Error en el procesamiento de los datos:** mala codificación en la grabación de los datos.
- **Errores de muestreo:** es inherente a trabajar con muestras en lugar de estudiar a toda la población. Hay que calcularlo, y éste dependerá del tamaño de la muestra y del tipo de muestreo utilizado. En el próximo apartado de análisis de los datos veremos cómo calcularlo.

7. Tasa de respuesta

7.1. Definiciones. Tasa de cooperación y de respuesta

Cuando se reporta la tasa de respuesta debe de quedar muy claro cómo se ha definido. De forma general hay que distinguir dos conceptos relacionados que muchas veces se confunden si no están bien especificados:

- **Tasa de cooperación:** es la proporción de todos los individuos entrevistados con entrevista completa, de todas las unidades elegibles con las que se ha contactado (23).

Personas con entrevista completa

Personas con entrevista completa + Personas con entrevista incompleta +
Personas que rechazan o no pueden hacer la entrevista

- **Tasa de respuesta:** es la proporción de todos los individuos entrevistados con entrevista completa de todas las unidades elegibles (23).

Personas con entrevista completa

Personas con entrevista completa + Personas con entrevista incompleta +
Personas que rechazan o no pueden hacer la entrevista +
No contacto (hogar o individuo elegible)

7.2. Tasa de respuesta aceptable

La tasa de respuesta de las encuestas de salud y estudios epidemiológicos en general, ha disminuido en las últimas décadas de forma generalizada (24). Por ejemplo, la participación en el NHIS en 1997 fue del 80,4 % en el cuestionario dirigido a adultos y del 84,1 % en la población infantil, mientras que en 2014 disminuyó al 58,9 % y 66,6 %, respectivamente (25). La tasa de respuesta del NHANES en 1999-2000 fue del 76,3 %, reduciéndose al 68,5 % en 2013-2014. En la encuesta de salud de Inglaterra, la tasa de respuesta en la muestra de adultos de 1994 fue del 71 % (26) y en 2017 disminuyó al 55 % (27). Lógicamente, esta respuesta es menor en el examen de salud. Por ejemplo, en la encuesta de salud de Inglaterra, solo el 27 % de los adultos aceptaron participar en 2017 en la extracción de una muestra de sangre (27). En España, comparando desde 2003 a 2016-2017, cuando se hace cargo el INE del diseño muestral y trabajo de campo (el diseño muestral de periodos anteriores no era comparable), la tasa de respuesta (entrevistas efectuadas respecto al total de hogares encuestables) descendió ligeramente del 75,6 % (28) al 69,9 % (13).

Las causas de esta disminución se relacionan tanto con la dificultad en localizar a las personas seleccionadas, como a una menor cooperación a participar una vez la muestra ha sido contactada, en consonancia con una disminución de la participación social (29).

En una encuesta de salud dirigida a población general una tasa de respuesta del 60 % podría ser aceptable (30). Sin embargo, la tasa de respuesta por sí sola no puede ser la única evidencia para juzgar la calidad de un estudio. Se necesita, además, conocer con detalle las características de los no participantes. Existe bastante consistencia de que las mujeres, las personas con alta posición socioeconómica o con mayor nivel educativo, casadas y con trabajo remunerado, participan más. Menos consistente es la relación con la edad, etnia o raza (31). Afortunadamente las características sociodemográficas pueden ser revertidas mediante la aplicación de ponderaciones. El problema se centra en el sesgo que pueda ocurrir en relación a las variables de exposición y de estado de salud, que no pueden solucionarse con un problema de ponderación. En este sentido, las personas que fuman, beben,

consumen drogas, suelen estar infrarrepresentadas (24). Además, los no respondedores suelen estar menos sanos en general, y como ha sido recientemente documentado en el NHIS, el riesgo de fallecer es mayor que los que aceptan participar en el estudio (31).

Por tanto, no es posible definir una tasa de respuesta aceptable. Es evidente que cuanto más alta, mejor, pero hay que tener en cuenta el análisis de la no respuesta para conocer con el mayor detalle posible los sesgos de selección de la muestra finalmente constituida.

7.3. Métodos para aumentar la tasa de respuesta

Existen varios procedimientos que permiten aumentar la tasa de respuesta. En primer lugar, valorar el método de entrevista. La tasa de respuesta más alta se obtiene con mucha diferencia con la entrevista personal cara a cara. Segundo, incrementar el número de intentos si no existe contacto con el hogar o la persona seleccionada, o no está disponible. Tercero, valorar dar incentivos (económicos o de otro tipo). Finalmente, comunicar con claridad los objetivos del estudio, en qué consiste la participación y la carga que implica. Una carta informativa, así como un contacto telefónico previos, pueden incentivar la participación en encuestas con entrevista personal cumplimentadas en el hogar. En entrevistas postales y online, además de los anteriores, la personalización de los cuestionarios, la calidad de la documentación, la reducción del cuestionario, la introducción de imágenes, uso de representaciones textuales en las categorías de respuesta, y asegurar la confidencialidad, pueden aumentar también la tasa de respuesta (32).

8. Análisis de datos con diseños de muestras complejas

El diseño muestral de una encuesta incluye habitualmente la necesidad de utilizar ponderaciones y correcciones en el cálculo del error estándar en el análisis de los datos.

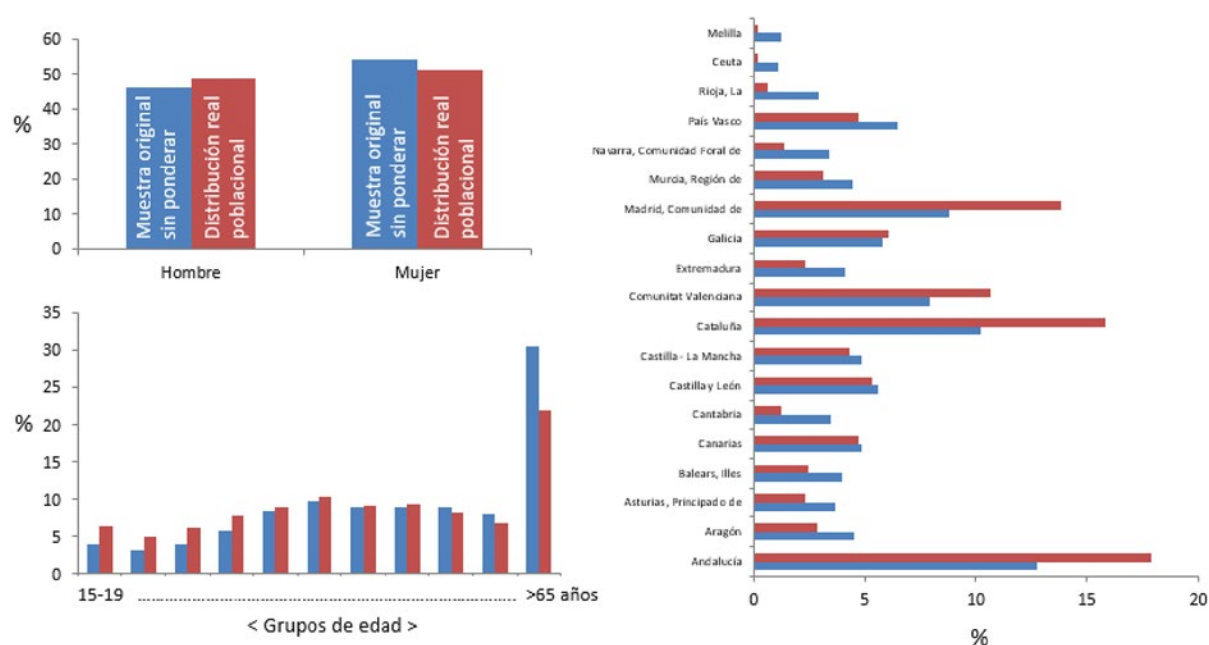
Las ponderaciones afectan al valor de las estimaciones y el diseño muestral complejo al error estándar, y, como consecuencia, a los intervalos de confianza.

Por ello, hay que estudiar detenidamente la metodología utilizada en las encuestas antes de diseñar el análisis de los datos.

8.1. Ponderación

Son necesarias para poder inferir a la población sin que estas estimaciones se desvíen de la distribución real de ciertas características sociodemográficas. Por una parte, algunas zonas geográficas están sobrerrepresentadas respecto a su distribución real en el conjunto de la población con el fin de poder hacer inferencias con un poder estadístico razonable. Por otra, al hacer el trabajo de campo no todos los estratos poblacionales participan de forma homogénea, por lo que se produce un sesgo de selección. Por ejemplo, en la figura 1 podemos apreciar cómo se distribuye la muestra de la ENSE 2016-17 (barras azules) según sexo, edad y comunidad autónoma de residencia, y qué diferencias hay con la distribución poblacional (barras rojas).

Figura 1. Ejemplo de estimaciones con y sin ponderación. Encuesta Nacional de Salud 2016-17.



Rápidamente observamos como en la ENSE responden más mujeres, participan más personas mayores y existen notables diferencias por comunidades autónomas. Por tanto, estas desviaciones hay que ajustarlas o calibrarlas mediante técnicas de ponderación. Simplemente hay que darle unos “pesos” o ponderaciones (multiplicar por unas fracciones) que rectifican estas desviaciones para que su distribución sea la distribución del conjunto de la población (barras azules).

Los criterios de ponderación de la ENSE 2016-17 se aplican a la siguiente multiestratificación generada por la combinación de las variables: Edad (grupos de 0-4, 5-9, 10-15, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65 y más); sexo; provincia; tamaño del hogar (1, 2, 3, 4 o más miembros). En total: $14 \times 2 \times 52 \times 5 = 7280$ valores diferentes.

Ahora vamos a describir cómo cambia el análisis al aplicar o no la ponderación, utilizando como ejemplo la pregunta sobre el estado de salud general. Este indicador es una proporción en la que describimos su distribución porcentual, realizando el análisis con el programa de análisis de datos Stata. Si no se utiliza ponderación, el 18,1 % de la muestra considera su estado de salud como muy bueno, mientras que el 2,2 % percibe que es muy malo. Al incluir la ponderación las estimaciones cambian a 21,3 % y 1,9 %.

```
. prop adulto_G21,percent cotype(norm)
```

Proportion estimation Number of obs = 23,089

	Percent	Std. Err.	Normal [95% Conf. Interval]	
adulto_G21				
Muy Bueno	18.15	0.25	17.65	18.64
Bueno	48.27	0.33	47.63	48.91
Regular	23.96	0.28	23.40	24.51
Malo	7.47	0.17	7.13	7.81
Muy malo	2.16	0.10	1.97	2.35

La pregunta adulto_G21 se refiere a cómo es su estado de salud

Sin ponderación

```
. prop adulto_G21 [pweight= adulto_FACTORADULTO],percent cotype(norm)
```

Proportion estimation Number of obs = 23,089

	Percent	Std. Err.	Normal [95% Conf. Interval]	
adulto_G21				
Muy Bueno	21.31	0.34	20.64	21.98
Bueno	49.20	0.40	48.42	49.98
Regular	21.19	0.31	20.57	21.80
Malo	6.38	0.18	6.02	6.74
Muy malo	1.92	0.10	1.71	2.13

Con ponderación

O en el siguiente ejemplo con la estimación de una variable continua, en este caso la edad. Si no se utiliza la ponderación la media de edad es 5 años mayor que si se ponderan los datos.

```
. mean adulto_EDADA
```

Mean estimation Number of obs = 23,089

	Mean	Std. Err.	[95% Conf. Interval]	
adulto_EDADA	53.43618	.1243481	53.19245	53.67991

La pregunta adulto_EDADA se refiere a la edad en años

Sin ponderación

```
. mean adulto_EDADA [pweight= adulto_FACTORADULTO]
```

Mean estimation Number of obs = 23,089

	Mean	Std. Err.	[95% Conf. Interval]	
adulto_EDADA	48.58008	.1509135	48.28428	48.87588

Con ponderación

Por tanto, las estimaciones puntuales: proporciones, prevalencias, medias, desviaciones estándar, medidas de asociación, etc., van a cambiar al ponderar los datos.

8.2. Cálculo del error estándar

Como hemos visto anteriormente, las encuestas poblacionales se basan en diseños muestrales complejos, introduciendo habitualmente procedimientos tanto de estratificación como de muestreo por conglomerados. Esto provoca que el cálculo de los errores estándar va a diferir respecto a si fueran estimados siguiendo un muestreo aleatorio simple.

La mayor parte de los grandes programas estadísticos incluyen módulos para el análisis de muestras complejas. Stata incorpora el módulo Survey Data [SVY].

Hay que configurar el “svyset”, señalando en primer lugar la variable que identifica al conglomerado, que en el caso de la ENSE es la sección censal. Posteriormente se configura la variable de ponderación, y finalmente la variable estrato, que se refiere a la división de los municipios en función de su tamaño. El comando de sintaxis:

1) `svyset IDconglomerado [pweight=IDponderacion], strata(IDestrato)`

Nombre de variable que identifica al conglomerado

Nombre de variable que identifica al factor de ponderación

Nombre de variable que identifica al estrato/s

2) Una vez definido el `svyset`, trabajamos con el comando “`svy:`”

```
. svyset seccion [pweight=adulto_FACTORADULTO], strata(Estrato)
      pweight: adulto_FACTORADULTO
      VCE: linearized
Single unit: missing
Strata 1: Estrato
SU 1: seccion
FPC 1: <zero>
```

Nombre de variable que identifica al conglomerado

Nombre de variable que identifica al factor de ponderación

Nombre de variable que identifica al estrato

`svy:prop adulto_G21,percent cotype(norm)`
(estimating proportion on estimation sample)

Survey: Proportion estimation

Number of strata = 19 Number of obs = 23,089
Number of PSUs = 2,498 Population size = 39,069,654
Design df = 2,479

	Percent	Linearized Std. Err.	Normal [95% Conf. Interval]	
adulto_G21				
Muy Bueno	21.31	0.42	20.49	22.13
Bueno	49.20	0.45	48.32	50.08
Regular	21.19	0.34	20.53	21.85
Malo	6.38	0.19	6.01	6.75
Muy malo	1.92	0.11	1.71	2.13

Con ponderación y SI corrección del error estándar

Comando (`svy:`) para trabajar análisis de datos con diseños muestrales complejos

`. prop adulto_G21 [pweight= adulto_FACTORADULTO], percent cotype(norm)`
Proportion estimation Number of obs = 23,089

	Percent	Std. Err.	Normal [95% Conf. Interval]	
adulto_G21				
Muy Bueno	21.31	0.34	20.64	21.98
Bueno	49.20	0.40	48.42	49.98
Regular	21.19	0.31	20.57	21.80
Malo	6.38	0.18	6.02	6.74
Muy malo	1.92	0.10	1.71	2.13

Con ponderación y NO corrección del error estándar

```
. svy:mean adulto_EDADa
(running mean on estimation sample)
```

Survey: Mean estimation

```
Number of strata =      19      Number of obs   =    23,089
Number of PSUs   =    2,498      Population size = 39,069,654
Design df        =              Design df      =    2,479
```

	Mean	Linearized Std. Err.	[95% Conf. Interval]	
adulto_EDADa	48.58008	.1650419	48.25645	48.90372

Con ponderación y SI corrección del error estándar

```
. mean adulto_EDADa [pweight= adulto_FACTORADULTO]
```

Mean estimation Number of obs = 23,089

	Mean	Std. Err.	[95% Conf. Interval]	
adulto_EDADa	48.58008	.1509135	48.28428	48.87588

Con ponderación y NO corrección del error estándar

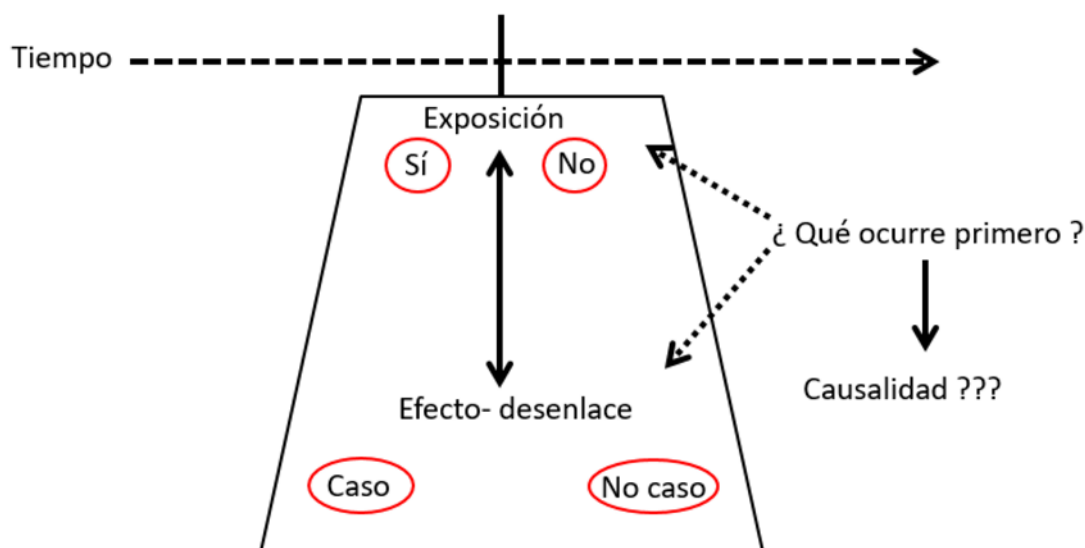
9. ¿Se pueden estudiar asociaciones?

Las encuestas de salud poblacionales tienen por lo general un diseño transversal, es decir, se recogen de forma simultánea en el tiempo información de la exposición y del desenlace (figura 2). Por tanto, no tenemos buena información temporal de la dirección de la asociación, y podemos tener un problema de bidireccionalidad (el factor de riesgo influye en el desenlace o el desenlace hace que se modifique el factor de riesgo). La utilidad de los estudios transversales analíticos para establecer asociación dependerá de la medida que la exposición actual refleje la exposición etiológicamente importante. Cuanto menos cambie la exposición desde el inicio, más útil será la medición actual.

Por ejemplo, si queremos en una encuesta transversal establecer asociaciones entre tabaco-depresión, obesidad-actividad física, o tener mascotas-asma, tendremos un gran problema de bidireccionalidad. Sin embargo, si la encuesta reconstruye exposiciones pasadas (aunque puede surgir otro problema como un error de medición debido al sesgo de memoria), se pueden aproximar asociaciones. Variables como la lactancia, exposición neonatal a humo de tabaco, el sexo, nivel educativo, etc. no son modificadas por el desenlace.

Por tanto, con diseños transversales sí podemos relacionar variables, describir su distribución, establecer correlaciones y plantear hipótesis. Pero no podemos realizar asociaciones causales.

Figura 2. No direccionalidad en los estudios transversales.



10. Encuestas de salud repetidas

Las encuestas de salud poblacionales suelen repetirse en el tiempo para monitorizar el estado de salud de la población. Lo ideal es que se repitan anualmente, porque permiten poder evaluar intervenciones, regulaciones y situaciones de salud que cambian debido a un factor externo de gran repercusión.

Sin embargo, hay que tener en cuenta que los estudios transversales, frente a los estudios longitudinales, confunden los efectos de edad y de cohorte, por lo que no diferencian si la causa de un cambio está en las diferencias de edad o en las diferencias en el momento del nacimiento (figura 3).

Figura 3. Confusión de los efectos de edad y de cohorte en los estudios transversales repetidos.

Grupo edad								
55-59	A	B	C	D	E	F	G	H
50-54	B	C	D	E	F	G	H	A
45-49	C	D	E	F	G	H	A	B
40-44	D	E	F	G	H	A	B	C
35-39	E	F	G	H	A	B	C	D
30-34	F	G	H	A	B	C	D	E
25-29	G	H	A	B	C	D	E	F
20-24	H	A	B	C	D	E	F	G
Año de estudio	1980	1985	1990	1995	2000	2005	2010	2015

11. Encuestas panel

A diferencia de los estudios transversales, este tipo de encuestas son estudios longitudinales donde se realiza un seguimiento a un mismo grupo de individuos (panel) por un periodo de tiempo, permitiendo analizar cómo cambian los individuos (u otras unidades de observación como los hogares) a lo largo del tiempo. Se basan en observaciones repetidas en el tiempo sobre los mismos temas en diferentes momentos.

En el campo de la epidemiología, las encuestas panel no se diferencian desde el punto de vista de diseño de un estudio de cohorte. Ambos son estudios longitudinales. Quizás, las mayores diferencias se relacionan con el objetivo del estudio: los diseños de cohortes investigan la asociación entre una o varias exposiciones y su efecto en la salud, mientras que las encuestas panel evalúan los cambios que se producen en el tiempo en las variables que son examinadas.

Un buen ejemplo de un estudio panel en el campo de la salud pública es la Encuesta sobre salud, envejecimiento y jubilación en Europa (SHARE) (33), una base de datos de panel multidisciplinar y transnacional sobre salud, situación socioeconómica y redes sociales y familiares, en la que participan 140 000 personas de 50 años o más (alrededor de 380 000 entrevistas). SHARE cubre 27 países europeos e Israel. Se han realizado hasta el momento 8 olas de entrevistas, iniciándose el estudio basal en 2004.

En España podemos destacar como estudios tipo panel la Encuesta de Condiciones de Vida (34). Se realiza en todos los países de la Unión Europea de forma anual, realizándose por primera vez en 2004. Las personas entrevistadas colaboran cuatro años seguidos. El objetivo principal de la encuesta es proporcionar información sobre la renta, el nivel y composición de la pobreza y la exclusión social en España y permitir la realización de comparaciones con otros países de la Unión Europea. Al proporcionar información longitudinal se puede conocer la evolución de las variables investigadas a lo largo del tiempo. Otros buenos ejemplos son la Encuesta de Presupuestos Familiares (35) o el Panel de Consumo Alimentario (36).

Entre las ventajas de las encuestas panel es que, desde el punto de vista descriptivo, permite diferenciar cambios en la distribución de la edad de cambios relacionados con un efecto cohorte. Además, desde una perspectiva temporal, el diseño longitudinal permite establecer relaciones causales. Entre las desventajas, destacar que son por lo general estudios más caros de realizar que los transversales, y se producen pérdidas en el seguimiento que pueden provocar un sesgo de selección y afectar a la representatividad de la muestra (37).

12. Referencias

1. Johnson TP, ed. *Handbook of Health Survey Methods*. Wiley; 2015.
2. Frees EW. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Repr. Univ. Press; 2008.
3. National Center for Health Statistics, Centers for Disease Control and Prevention. NCHS Data Linked to NDI Mortality Files. <https://www.cdc.gov/nchs/data-linkage/mortality.htm>
4. Li Y, Schoufour J, Wang DD, *et al.* Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. *BMJ*. 2020;l6669. doi:[10.1136/bmj.l6669](https://doi.org/10.1136/bmj.l6669)
5. Knott CS, Coombs N, Stamatakis E, Biddulph JP. All cause mortality and the case for age specific alcohol consumption guidelines: pooled analyses of up to 10 population based cohorts. *BMJ*. 2015;350:h384. doi:[10.1136/bmj.h384](https://doi.org/10.1136/bmj.h384).
6. Powell-Griner E, Madans J. History of the National Health Interview Survey. *Proc Am Stat Assoc Section on Health Policy Statistics*. 2007.

7. National Center for Health Statistics, Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey. <https://www.cdc.gov/nchs/nhanes/index.htm>
8. *Health Survey for England*. NatCen Social Research; 2018.
9. Ministerio de Sanidad. Encuesta Nacional de Salud de España. <https://www.mscbs.gob.es/estadEstudios/estadisticas/encuestaNacional/home.htm>
10. Ministerio de Sanidad. Encuesta Europea de Salud en España 2014. https://www.mscbs.gob.es/estadEstudios/estadisticas/EncuestaEuropea/Enc_Eur_Salud_en_Esp_2014.htm
11. Taherdoost H. Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. SSRN Electronic Journal. doi:[10.2139/ssrn.3205035](https://doi.org/10.2139/ssrn.3205035).
12. Pastor-Barriuso R. Bioestadística. Escuela Nacional de Sanidad; Centro Nacional de Epidemiología; 2012. <http://gesdoc.isciii.es/gesdoccontroller?action=download&id=03/06/2013-7dd67975c5>
13. Ministerio de Sanidad. Encuesta Nacional de Salud 2017. Metodología. https://www.mscbs.gob.es/estadEstudios/estadisticas/encuestaNacional/encuestaNac2017/ENSE17_Metodologia.pdf
14. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System. <https://www.cdc.gov/brfss/index.html>
15. Consejería de Sanidad. Comunidad de Madrid. Sistema de Vigilancia de Factores de Riesgo de Enfermedades Crónicas (SIVFRENT). <https://www.comunidad.madrid/servicios/salud/factores-riesgo-enfermedad>
16. Instituto Nacional de Estadística. Encuesta sobre equipamientos y uso de tecnología de la información y comunicación en los hogares. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=estadistica_C&cid=1254736176741&menu=ultiDatos&idp=1254735976608
17. Galán I, Rodríguez-Artalejo F, Zorrilla B. Comparación entre encuestas telefónicas y encuestas «cara a cara» domiciliarias en la estimación de hábitos de salud y prácticas preventivas. *Gac Sanit*. 2004;18(6):440-450. doi:[10.1016/S0213-9111\(04\)72031-2](https://doi.org/10.1016/S0213-9111(04)72031-2)
18. Universidad de Sevilla. Estudio Health Behaviour in School-aged Children (HBSC). <https://www.hbsc.es/>
19. Stern MJ, Bilgen I, Dillman DA. The State of Survey Methodology: Challenges, Dilemmas, and New Frontiers in the Era of the Tailored Design. *Field Methods*. 2014;26(3):284-301. doi:[10.1177/1525822X13519561](https://doi.org/10.1177/1525822X13519561)

20. Connor Gorber S, Tremblay M, Moher D, Gorber B. A comparison of direct vs. self-report measures for assessing height, weight and body mass index: a systematic review. *Obes Rev.* 2007;8(4):307-326. doi:[10.1111/j.1467-789X.2007.00347.x](https://doi.org/10.1111/j.1467-789X.2007.00347.x).
21. Mackenbach JP, Simon JG, Looman CWN, Joung IMA. Self-assessed health and mortality: could psychosocial factors explain the association? *Int J Epidemiol.* 2002;31(6):1162-1168. doi:[10.1093/ije/31.6.1162](https://doi.org/10.1093/ije/31.6.1162).
22. Sánchez-López M del P, Dresch V. The 12-Item General Health Questionnaire (GHQ-12): reliability, external validity and factor structure in the Spanish population. *Psicothema.* 2008;20(4):839-843.
23. American Association for Public Opinion Research. Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys. Revised 2016. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf
24. Galea S, Tracy M. Participation Rates in Epidemiologic Studies. *Ann Epidemiol.* 2007;17(9):643-653. doi:[10.1016/j.annepidem.2007.03.013](https://doi.org/10.1016/j.annepidem.2007.03.013).
25. Czajka J, Beyler A. Declining Response Rates in Federal Surveys: Trends and Implications. <https://www.mathematica.org/our-publications-and-findings/publications/declining-response-rates-in-federal-surveys-trends-and-implications-background-paper>
26. Mindell J, Biddulph JP, Hirani V, *et al.* Cohort Profile: The Health Survey for England. *Int J Epidemiol.* 2012;41(6):1585-1593. doi:[10.1093/ije/dyr199](https://doi.org/10.1093/ije/dyr199).
27. NHS Digital. Health Survey for England 2017. Methods. <https://files.digital.nhs.uk/D4/EFB4CD/HSE17-Methods-rep.pdf>
28. Ministerio de Sanidad. Evaluación de la falta de respuesta en la Encuesta Nacional de Salud 2003. https://www.mscbs.gob.es/estadEstudios/estadisticas/docs/evalfrsalud_definitivo.pdf
29. Morton LM. Encouraging participation in medical research: what strategies work? *J Clin Epidemiol.* 2008;61(10):969-970. doi:[10.1016/j.jclinepi.2007.09.001](https://doi.org/10.1016/j.jclinepi.2007.09.001).
30. Mundy D. A question of response rate. *Science Editor.* 2002;25 (1); Solution corner section, para.2.
31. Keyes KM, Rutherford C, Popham F, Martins SS, Gray L. How healthy are survey respondents compared with the general population?: Using Survey-linked Death Records to Compare Mortality Outcomes. *Epidemiology.* 2018;29(2):299-307. doi:[10.1097/EDE.0000000000000775](https://doi.org/10.1097/EDE.0000000000000775).

32. Edwards PJ, Roberts I, Clarke MJ, *et al.* Methods to increase response to postal and electronic questionnaires. Cochrane Methodology Review Group, ed. *Cochrane Database Syst Rev*. Published online July 8, 2009. doi:[10.1002/14651858.MR000008.pub4](https://doi.org/10.1002/14651858.MR000008.pub4).
33. SHARE - Survey of Health, Ageing and Retirement in Europe. <https://share-eric.eu>
34. Instituto Nacional de Estadística. Encuesta de condiciones de vida. Metodología. https://www.ine.es/daco/daco42/condivi/ecv_metodo.pdf
35. Instituto Nacional de Estadística. Encuesta de Presupuestos Familiares. Metodología. <https://www.ine.es/metodologia/t25/t2530p458.pdf>
36. Ministerio de Agricultura, Pesca y Alimentación. Panel de Consumo Alimentario. <https://www.mapa.gob.es/es/alimentacion/temas/consumo-tendencias/panel-de-consumo-alimentario/>
37. Lynn P. Metodología de las encuestas longitudinales. 2005. https://www.eustat.eus/productosServicios/datos/Sem45_c.pdf

Capítulo 9. Estudios de intervención

MIGUEL ÁNGEL ROYO BORDONADA

1. Introducción
2. Asociación vs. causalidad
3. Ensayos clínicos vs. ensayos preventivos
4. Planificación de un estudio de intervención. El protocolo
5. Aleatorización
6. Diseño del ensayo
7. Los participantes
8. La intervención
9. Enmascaramiento
10. Análisis de un ensayo
11. Sesgos en los ensayos
12. La declaración CONSORT
13. Conclusión
14. Referencias

Anexo I. Lista de comprobación de la declaración SPIRIT 2013: elementos que se recomienda abordar en el protocolo de un ensayo clínico y en los documentos relacionados.

Anexo II. Lista de verificación TIDieR (plantilla para descripción de intervención y replicación)*: información a incluir al describir una intervención y la ubicación de la información.

Anexo III. Consort 2010 lista de comprobación de la información que hay que incluir al comunicar un ensayo clínico aleatorizado*

1. Introducción

La Real Academia de la Lengua Española establece que un experimento consiste en hacer operaciones destinadas a descubrir, comprobar o demostrar determinados fenómenos o principios científicos. (1) Para la mayoría de los científicos, esta definición abarca cualquier conjunto de observaciones, realizadas bajo condiciones controladas, que permitan contrastar una hipótesis científica previamente planteada. Por ejemplo, Otto von Guericke (1602-1686) propuso la teoría del giro de los cometas alrededor del sol en una órbita elíptica excéntrica. Un año después de su muerte, Isaac Newton estableció la Ley de Gravitación Universal. Posteriormente, Edmund Halley, que había descubierto el cometa que lleva su nombre en 1682, se propuso comprobar la hipótesis de Otto. A partir de las leyes de Newton, calculó que el Halley recorría su órbita en setenta y seis años, por lo que supuso que era el mismo cometa que se había visto en ocasiones anteriores (1607 y 1531) y postuló que aparecería de nuevo en 1758. Lamentablemente, Halley no vivió lo suficiente para verlo. Quien sí lo hizo fue Johann Georg Palitzsch, un terrateniente alemán aficionado a la astronomía que estuvo todo un año observando el cielo a la espera del cometa. Por fin, el 25 de diciembre de 1758 vio la estela del cometa Halley en el firmamento, confirmando la hipótesis de Otto. (2) Las personas involucradas en esta investigación se limitaron a observar un fenómeno y analizar los datos recogidos acerca del mismo, pero sin intervenir en su desarrollo, que siguió su curso natural. Desde esta perspectiva, los estudios de cohortes y los de casos y controles también serían experimentos, en el sentido más amplio del término. Sin embargo, en epidemiología distinguimos entre estudios descriptivos, que permiten generar hipótesis a partir de la descripción de los fenómenos observados, analíticos observacionales, que permiten contrastar hipótesis mediante la observación de los fenómenos tal cual se producen, aun cuando el mero hecho de observar y recoger datos puede influir en los hechos observados, y experimentales, donde el investigador manipula de forma consciente las condiciones del estudio, con el objeto de averiguar el efecto de tal manipulación (intervención) sobre los fenómenos observados.

Puesto que el elemento distintivo de un estudio experimental es la aplicación por el investigador de una intervención que altera las condiciones del mismo y cuyo efecto se desea analizar, nos referiremos a los mismos como estudios de intervención. Para que podamos hablar con propiedad de un estudio de intervención, deben cumplirse las siguientes condiciones:

1. La única razón por la que los sujetos reciben la intervención es el cumplimiento del protocolo del estudio.

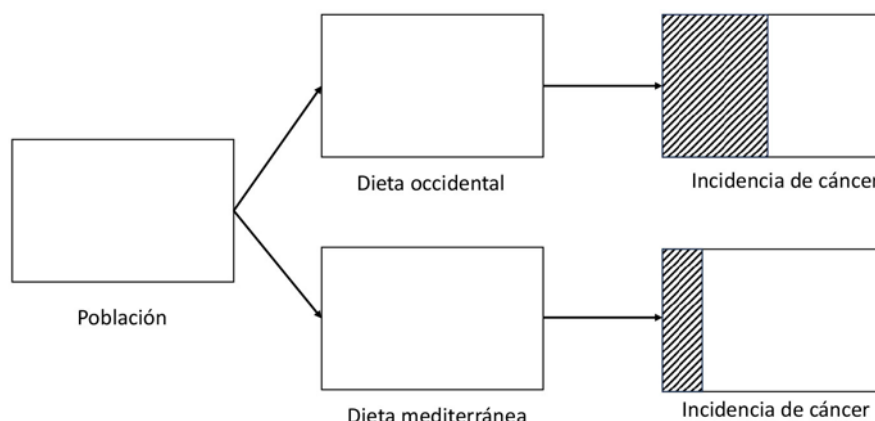
2. Existe una serie de sujetos (grupo control) que no reciben la intervención (experimento controlado).
3. La asignación de la intervención se lleva a cabo mediante un mecanismo que depende del azar (experimento aleatorizado).

Los estudios de intervención (experimentos controlados y aleatorizados) se denominan comúnmente ensayos. Cuando no se dan las condiciones previas, bien porque la intervención se ha producido de forma espontánea (experimentos naturales), no existe un grupo control (experimentos no controlados) o la intervención ha sido asignada mediante un mecanismo distinto al azar (experimentos controlados sin aleatorizar), hablamos de pseudoexperimentos o estudios cuasiexperimentales, aunque a efectos de análisis e interpretación de los mismos deberán ser tratados como estudios observacionales.

2. Asociación vs. causalidad

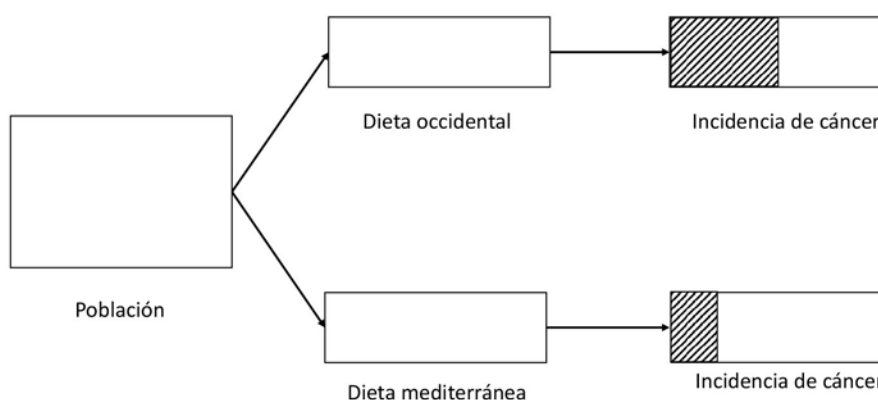
Imagine que seguimos a una población que sigue una dieta occidental durante diez años y medimos la incidencia de cáncer. Si pudiéramos retroceder en el tiempo hasta la fecha en la que se inició el seguimiento, administrar una intervención preventiva a todo el grupo (por ejemplo, pasarse a una dieta mediterránea) y repetir el proceso de seguimiento y medición, exactamente igual que la primera vez, cualquier cambio en la incidencia de cáncer podría atribuirse a la intervención (Figura 1). (3) Y esto es así porque la única diferencia entre ambos grupos consistiría en la intervención administrada; por lo demás serían idénticos, de hecho, se trataría de las mismas personas sujetas a las mismas experiencias vitales durante todo el período de seguimiento.

Figura 1. Efecto de pasarse a una dieta mediterránea en la incidencia de cáncer (situación hipotética, experimental).



En el ejemplo hipotético previo, la asociación observada entre la intervención y la incidencia de cáncer es causal. Por el contrario, en el mundo real (Figura 2) las personas que siguen una dieta mediterránea por iniciativa propia son distintas y están sujetas a experiencias diferentes de las que siguen una dieta occidental, todo lo cual puede influir en la incidencia de cáncer. Por tanto, en condiciones observacionales, asociación no es causalidad (Figura 2), ya que no podemos descartar que la asociación observada se deba, no a la exposición en estudio, sino a otros factores desconocidos o mal controlados, sean estos personales (genéticos, sociodemográficos) o ambientales (entornos físico, social, cultural o político).

Figura 2. Efecto de seguir una dieta mediterránea en la incidencia de cáncer (situación real, observacional).



Como no es posible viajar en el tiempo, lo que hacemos es comparar dos o más grupos de personas, uno de los cuales recibe la intervención bajo estudio. Pues bien, para poder establecer relaciones causales en el mundo real necesitamos crear unas

condiciones tales que el único elemento de variación entre los grupos comparados sea la intervención de interés. Pero los factores que influyen en la salud son múltiples, con frecuencia desconocidos, y presentan complejas interacciones, por lo que resulta materialmente imposible crear un conjunto de condiciones que conduzcan invariablemente al desarrollo de una enfermedad concreta en un determinado periodo de tiempo, incluso cuando la población de estudio sea un grupo de ratones clonados en el laboratorio. Inevitablemente, siempre existirán variaciones biológicas en los factores que contribuyen al desarrollo de la enfermedad en cuestión. Aunque consigamos mantener idénticas condiciones de temperatura, humedad, hábitat e iluminación, lo cual es poco probable, no podemos evitar las diferencias individuales en el metabolismo basal ni hacer que todos los animales coman exactamente lo mismo o realicen ejercicio físico de igual intensidad y duración. Con los seres humanos todo se complica, al entrar en juego factores socioeconómicos y culturales sobre los que el investigador carece de control. En resumidas cuentas, la idea de crear dos grupos perfectamente equiparables, salvo por la intervención de interés, no es realista. No obstante, puede afirmarse que los grupos son intercambiables a efectos prácticos cuando la variación de los factores implicados en la condición bajo estudio es tan pequeña que su posible influencia es despreciable en comparación con el previsible efecto de la intervención de interés. Pues bien, los estudios de intervención, a diferencia de los observacionales, permiten asegurar la intercambiabilidad de los grupos comparados cuando se cumple lo siguiente: (3)

1. Asignación de los participantes a uno u otro grupo mediante un mecanismo aleatorio que garantice el equilibrio de los factores pronósticos entre los grupos comparados.
2. Ausencia de interferencia o contaminación: el resultado de un sujeto del estudio no depende de la intervención aplicada a otros participantes. Existe mayor riesgo de interferencia en estudios sobre agentes infecciosos o con intervenciones educativas, por interacción social entre los participantes.
3. Homogeneidad: existe una única versión de la intervención. Las posibles variaciones a la hora de aplicar la intervención por distintos profesionales, en momentos o lugares diferentes, deben ser irrelevantes para su eficacia. Puede no darse cuando el resultado de la intervención depende de la habilidad o conocimientos del profesional (intervenciones quirúrgicas, educativas o cognitivo-conductuales) o de los recursos disponibles para aplicarla.
4. Adherencia al protocolo. La falta de adherencia puede deberse al abandono de la intervención asignada, su aplicación incorrecta o la administración de intervenciones no previstas en el protocolo.

5. Ausencia de pérdidas al seguimiento. Cuando un individuo se pierde, resulta difícil o imposible averiguar el desenlace del mismo en relación con la condición bajo estudio.
6. Aplicación de algún tipo de intervención en el grupo control (práctica habitual, intervención inactiva, intervención simulada, intervención parcial, etc.).

En un estudio de intervención donde se dan las condiciones previas, asociación es causalidad. Los componentes básicos de un ensayo para garantizar la intercambiabilidad de los grupos comparados son la aleatorización, el enmascaramiento y una planificación cuidadosa del mismo, que incluya la estandarización de todos los procedimientos previstos y, en particular, de la intervención que se administrará.

3. Ensayos clínicos vs. ensayos preventivos

Los ensayos clínicos evalúan el efecto de una intervención, sea esta preventiva (prevención secundaria) o terapéutica (curativa o rehabilitadora), en el pronóstico de una enfermedad concreta; mientras que los ensayos preventivos buscan reducir la aparición de factores de riesgo (prevención primordial) o enfermedades (prevención primaria) en personas sanas. Aunque los principios metodológicos de ambos tipos de ensayos son los mismos, existen ciertas consideraciones a tener en cuenta en el diseño, planificación y organización de los ensayos preventivos.

3.1. Ensayos clínicos

En 1935 Sir Austin Bradford Hill formuló los principios de los ensayos clínicos. (4,5) Un ensayo clínico es un experimento planificado en el que, de forma prospectiva, se comparan dos o más intervenciones, o intervención y placebo, asignadas de forma aleatoria a una muestra de pacientes para evaluar su eficacia y seguridad. (6) La selección de los sujetos y los períodos de tratamiento y seguimiento tienen lugar simultáneamente en los grupos comparados. El primer ensayo clínico con un grupo control aleatorizado se realizó en 1948 para evaluar la eficacia de la estreptomina en el tratamiento de la tuberculosis pulmonar. (7) Los ensayos clínicos con intervenciones preventivas (prevención secundaria) tienen por objeto el diagnóstico y tratamiento precoz de la enfermedad, antes de que esta produzca síntomas, o la prevención (evitar o retrasar) de sus recidivas y secuelas, incluyendo

la muerte, por ej., en el ensayo ISIS-2, una parte de los pacientes recibieron aspirina tras sufrir un infarto de miocardio para comprobar si este fármaco reducía la incidencia de nuevos accidentes cardiovasculares. (8)

En los ensayos clínicos para evaluar el efecto de un nuevo tratamiento, el uso de placebo en el grupo control solo es aceptable si no existe un tratamiento alternativo para la enfermedad en cuestión o, en caso de existir, si el retraso en la aplicación del mismo no conlleva riesgo de efectos adversos permanentes y los participantes son informados de las alternativas disponibles. (9) Tradicionalmente, los ensayos clínicos han evaluado el efecto de las intervenciones en condiciones ideales (eficacia), con participantes sometidos a una estrecha vigilancia. Además, los ensayos no incluyen todo el espectro de posibles pacientes y los que aceptan participar suelen estar más concienciados con su estado de salud y mejor motivados para el cambio, lo que favorece una mayor adherencia al tratamiento. Estos ensayos se denominan explicativos; mientras que los que evalúan el efecto de la intervención en condiciones muy similares a las de la práctica clínica real (efectividad) se denominan ensayos pragmáticos. (10) Los ensayos explicativos, en su afán por demostrar la eficacia de la intervención, a veces incluyen solo a participantes que han mostrado previamente un buen nivel de adherencia a la misma.

3.2. Ensayos preventivos

Los ensayos preventivos o ensayos de campo presentan una mayor complejidad logística, pues al trabajar con personas sanas, que no buscan atención sanitaria, hay que realizar un importante trabajo de campo para su reclutamiento y establecer centros donde llevar a cabo el estudio. Además, requieren una muestra mayor para conseguir que el desenlace de interés se produzca en un número suficiente de participantes. En el mayor ensayo preventivo llevado a cabo hasta ahora en humanos, realizado en 1954, se administró vacuna Salk o placebo a 800 000 escolares estadounidenses. (11) A pesar de la baja incidencia de la poliomielitis (1/2000 anual), la gravedad de la misma justificó el esfuerzo y los recursos que fue necesario invertir para realizar el ensayo. Los resultados fueron concluyentes: la incidencia de poliomielitis y enfermedad parálitica se redujeron a la mitad en el grupo de intervención, y las cuatro muertes por poliomielitis se produjeron en el grupo control.

En un ensayo clínico, la aplicación de la intervención viene determinada por el diagnóstico de la enfermedad; mientras que en un ensayo preventivo no siempre está claro cuál es el mejor momento para intervenir (ventana de exposición), ya que el proceso de génesis de muchas enfermedades no transmisibles, como las vasculares y el cáncer, puede comenzar en épocas muy precoces de la vida y prolon-

garse durante décadas. (12) En la práctica, la necesidad de disponer de un número de casos suficiente para contrastar la hipótesis del estudio en un período de tiempo limitado lleva a seleccionar sujetos de alto riesgo: personas de edad avanzada, próxima al momento en que suele aparecer la enfermedad de interés, y expuestas a factores de riesgo de la misma. Por ejemplo, en el *Alpha Tocopherol Beta Carotene Cancer Prevention Study* (en adelante, ATBC), un ensayo controlado con placebo para determinar si los suplementos de α -tocoferol y β -caroteno podían prevenir el cáncer de pulmón, se seleccionaron varones fumadores de 50 a 69 años. (13) Contrariamente a lo esperado, la incidencia de cáncer de pulmón no se redujo con los suplementos de α -tocoferol y aumentó un 18 % con los suplementos de β -caroteno. La hipótesis del ensayo estaba respaldada por los resultados de estudios epidemiológicos observacionales, los cuales habían encontrado un menor riesgo de cáncer con niveles séricos y de ingesta dietética de β -caroteno, procedente de frutas y verduras, por encima de la media. (14) Es posible que el β -caroteno tenga un efecto preventivo únicamente en fases precoces de la carcinogénesis y no en estadios más avanzados del proceso, con la posible existencia de lesiones precancerosas, de modo que el ensayo habría tenido lugar pasado el periodo de tiempo durante el cual el sujeto es susceptible al efecto de la intervención (ventana de exposición). Además, la suplementación con β -caroteno es una simplificación de una intervención con frutas y verduras: es posible que sea un marcador de su consumo, pero sin influencia en la prevención del cáncer, que sería atribuible a otros componentes protectores de las frutas y verduras. Y aunque el β -caroteno sea uno de los componentes activos en la prevención del cáncer, también es posible que solo pueda ejercer su efecto cuando se consume junto con otros nutrientes o sustancias quimiopreventivas, tal como se encuentran combinadas en los alimentos. Por último, también cabe la posibilidad de que el β -caroteno prevenga el cáncer a las dosis alcanzadas con una ingesta alimentaria variada, mientras que a dosis varias veces más altas, como las alcanzadas en este ensayo, posea efectos tóxicos, lo cual explicaría el aumento del riesgo de cáncer entre los consumidores de suplementos.

En comparación con los ensayos clínicos, los ensayos preventivos tienden en mayor medida a un enfoque pragmático (evaluación de la efectividad). Sin embargo, no existen ensayos puramente pragmáticos o puramente explicativos, sino más o menos pragmáticos, en función de cuanto se aproxime el diseño del ensayo a las condiciones reales de aplicación de la intervención. (15)

4. Planificación de un estudio de intervención. El protocolo

La planificación cuidadosa de un ensayo es fundamental para el correcto desarrollo del mismo. El plan del ensayo debe plasmarse en un protocolo por escrito, donde se justifique la oportunidad del mismo y se establezcan los procedimientos para su conducción, supervisión y evaluación. En respuesta a las lagunas o deficiencias en muchos protocolos de ensayos clínicos y la variación en cuanto al alcance y las recomendaciones de las guías para la elaboración de los mismos, (16) un grupo de más de un centenar de expertos (investigadores, profesionales de la salud y la ética, metodólogos, estadísticos, coordinadores de ensayos, editores de revistas biomédicas, financiadores y representantes de organismos reguladores) elaboraron la declaración SPIRIT (*Standard Protocol Items: Recommendations for Interventional Trials*): Ítems estándar de un protocolo: recomendaciones para los ensayos de intervención. (17) Esta declaración define un protocolo como «el documento que proporciona suficientes detalles a fin de: 1) entender los antecedentes, la justificación, los objetivos, la población de estudio, las intervenciones, los métodos, los análisis estadísticos, las consideraciones éticas, los planes de diseminación y la administración del ensayo; 2) reproducir los aspectos clave de los métodos y la conducción del ensayo; y 3) evaluar el rigor científico y ético del ensayo desde su aprobación hasta la diseminación de los resultados».

La declaración SPIRIT proporciona una guía sobre el contenido mínimo que debe tener el protocolo de un ensayo, con un listado de 33 ítems agrupados en los siguientes apartados:

1. **Información administrativa:** título, filiación y funciones de los responsables del mismo, y datos relativos al registro del ensayo. Asimismo, recomienda incluir el conjunto de datos del registro de ensayos clínicos de la Organización Mundial de la Salud.
2. **Introducción:** antecedentes y justificación, objetivos y diseño.
3. **Métodos:** participantes, intervenciones, resultados, asignación de las intervenciones (aleatorización), recolección, gestión y análisis de datos, y monitorización del ensayo.
4. **Ética y diseminación:** todo ensayo requiere de la aprobación previa de un comité de ética de la investigación, que tendrá en cuenta aspectos relacionados con el consentimiento informado, la confidencialidad, las declaraciones de intereses de los implicados y el acceso a los datos. También se establecerán los planes para comunicar posibles enmiendas al protocolo, la política de diseminación de los resultados y, en caso necesario, las disposiciones para atender a los participantes tras finalizar el ensayo.

5. **Apéndices:** Incluye el modelo del consentimiento informado y los planes para recoger y manejar muestras biológicas, cuando proceda. En ocasiones, algunos detalles relevantes pueden aparecer en documentos relacionados, como el plan de análisis estadístico, el formulario de recogida de datos, el manual de operaciones o el contrato con el investigador.

En el anexo 1 se presenta la lista completa de ítems, con una breve descripción de los mismos. Se recomienda que la lista de comprobación se lea junto con el documento *SPIRIT 2013 Explanation and Elaboration*, donde se proporciona información detallada y aclaraciones importantes sobre cada ítem. (18)

Por último, al planificar un ensayo es preciso atenerse a la legislación aplicable en cada momento y lugar. En España, la investigación clínica con medicamentos de uso humano para el tratamiento y prevención de enfermedades está sujeta al Real Decreto 1090/2015, de 4 de diciembre, por el que se regulan los ensayos clínicos con medicamentos, los Comités de Ética de la Investigación con medicamentos y el Registro Español de Estudios Clínicos. (19) Los ensayos con intervenciones no farmacológicas están sujetos a la misma normativa en todo aquello que resulte de aplicación a los mismos. El Decreto contempla la modalidad de ensayo clínico de bajo nivel de intervención, aplicable cuando se dispone de pruebas sobre la seguridad y eficacia de los medicamentos en investigación y los procedimientos complementarios de diagnóstico o seguimiento entrañan un riesgo o carga adicional para la seguridad de los sujetos que es mínimo comparado con el de la práctica clínica habitual. Desde esta perspectiva, la mayoría de los ensayos de prevención con intervenciones sobre estilos de vida se considerarían de bajo nivel de intervención.

En los siguientes apartados del capítulo profundizaremos en los aspectos metodológicos fundamentales en la planificación y ejecución de un ensayo, a saber: aleatorización, diseño del ensayo, intervención, participantes, enmascaramiento, análisis y posibles sesgos.

5. Aleatorización

La aleatorización consiste en la asignación de los participantes de un ensayo a los diferentes grupos de intervención utilizando un mecanismo aleatorio, de forma que ni el investigador encargado del reclutamiento ni los futuros participantes conozcan la intervención que va a ser asignada en el momento de la inclusión de estos en el estudio (ocultamiento de la secuencia de asignación), una vez obtenido

su consentimiento informado. (20) El objetivo de la aleatorización es producir grupos similares en tamaño e intercambiables, con una distribución homogénea de las características conocidas y desconocidas que pudieran afectar al desenlace de interés (factores de confusión), de modo que cualquier diferencia en la distribución de estos factores entre los grupos comparados se deba exclusivamente al azar.

5.1. Ocultamiento de la secuencia de asignación

Aunque el mecanismo de generación de la secuencia de asignación sea aleatorio, si quienes reclutan a los participantes saben a qué intervención será asignado el próximo sujeto que se incluya en el estudio, porque tienen acceso a la secuencia de asignación, el sesgo de selección está servido. Este sesgo se produce cuando el encargado del reclutamiento, guiado —consciente o inconscientemente— por sus preferencias personales, rechaza a un participante que hubiera preferido ver asignado a otra intervención, en base a sus factores pronósticos, o retrasa su inclusión en el estudio hasta que salga la intervención deseada. Por ejemplo, en un ensayo donde se pretendía comparar el tratamiento habitual para el dolor de espalda (terapia tradicional) con un nuevo tratamiento (terapia activa), los sujetos asignados a la terapia activa fueron más numerosos (165 vs. 66) y presentaban un dolor más leve que los que recibieron el tratamiento habitual. (21) El ocultamiento de la secuencia de asignación evita un sesgo de selección en la aleatorización y no debe confundirse con el enmascaramiento de la intervención asignada (a los participantes, los que la administran y los que miden la variable respuesta), que previene de sesgos por falta de adherencia y en la medición del desenlace de interés. Además, a diferencia del ocultamiento de la secuencia de asignación, el enmascaramiento no siempre es factible (intervenciones sobre estilos de vida) o su aplicación presenta dificultades prácticas o éticas (intervenciones quirúrgicas).

El método más seguro de ocultamiento de la secuencia de asignación consiste en utilizar un sistema de asignación aleatoria remoto y centralizado. Una vez reclutado un nuevo participante para el ensayo e incorporados sus datos al sistema informático, este lo asigna de forma automática a un grupo de intervención y lo notifica al encargado del reclutamiento mediante un mensaje por teléfono o internet. En caso de utilizar sobres para la asignación, estos deben ser opacos, numerados secuencialmente y sellados, abriéndose únicamente en el momento de la asignación de cada nuevo participante.

5.2. Métodos de aleatorización

Los métodos de aleatorización se dividen en dos tipos. Por un lado, los métodos de asignación fija, en los que la probabilidad de ser asignado a uno u otro grupo es constante a lo largo del período de inclusión de sujetos en el ensayo. Por otro lado, los métodos de asignación adaptativa, en los que esa probabilidad varía en el tiempo en función de la respuesta ante la intervención en los sujetos previamente incluidos en el ensayo o de la existencia de desequilibrios en el número de participantes o en la distribución de los factores pronósticos entre los grupos comparados. (6)

5.2.1. Aleatorización simple

Dentro de los métodos de asignación fija, la aleatorización simple constituye el procedimiento más elemental, intuitivo y fácil de aplicar. Si tenemos dos grupos de comparación, podemos lanzar una moneda al aire y asignar a los sujetos a uno u otro grupo en función de si sale cara o cruz. El mismo método puede llevarse a cabo de forma menos rudimentaria con una tabla de números aleatorios o, con mayor nivel de sofisticación, mediante la generación de números pseudoaleatorios por medio de un programa informático, una técnica muy conveniente, que se puede aplicar con facilidad de forma centralizada en estudios multicéntricos.

Cuando el tamaño de la muestra es pequeño, la aleatorización simple no garantiza el equilibrio en el tamaño de los grupos ni la distribución homogénea de los factores de confusión. A modo de orientación, por debajo de 100 sujetos por grupo de comparación es muy posible que la aleatorización no consiga su propósito; mientras que por encima de 300 sujetos por grupo se puede asumir la intercambiabilidad. (22) Por ello, al presentar los resultados de un ensayo, lo primero que hay que mostrar es una tabla con la distribución de los factores pronósticos conocidos en los grupos comparados. El equilibrio en el tamaño de los grupos y una distribución homogénea de los factores pronósticos constituyen la mejor prueba de que la aleatorización ha funcionado y nos indica que los grupos también serán comparables respecto de factores pronósticos desconocidos, lo cual, en último término, nos permitirá atribuir cualquier diferencia observada en el desenlace de interés a la intervención investigada.

5.2.2. Aleatorización por bloques

Cuando el número de sujetos del ensayo no es lo bastante grande para que la aleatorización simple garantice el equilibrio en el tamaño de los grupos (tamaños desiguales disminuyen la potencia de los test estadísticos), se usa la aleatorización por bloques de pequeño tamaño. Las intervenciones se asignan aleatoriamente

dentro de cada bloque, con la salvedad de que el número de sujetos asignados a cada intervención debe ser el mismo. Por ejemplo, en un ensayo clínico con dos grupos (tratamiento y placebo) y bloques de ocho pacientes, en cada bloque cuatro recibirán el tratamiento y cuatro el placebo. Supongamos que lanzamos la moneda al aire (cara/tratamiento y cruz/placebo) y obtenemos cara. Continuamos lanzando la moneda y por azar obtenemos cuatro cruces seguidas. A partir de ese momento ya no volveríamos a lanzar la moneda, sino que asignaríamos «por decreto» a los tres pacientes restantes del bloque al tratamiento. El problema con este método es que el investigador podría adivinar con antelación el grupo al que van a ser asignados los tres últimos pacientes del bloque, con el consiguiente riesgo de sesgo de selección. Una forma de evitar este problema consiste en hacer bloques de distinto tamaño cuyo orden también se asigna de forma aleatoria.

5.2.3. Aleatorización estratificada

El problema derivado del posible desequilibrio en los factores pronósticos puede ser abordado a través de técnicas analíticas de ajuste o mediante aleatorización estratificada. El número de estratos es el producto de las categorías de los factores considerados, por ej., si estos son sexo (hombre o mujer) y estado civil (soltero, casado, separado o viudo), tendremos ocho estratos. Cuando un nuevo sujeto es incorporado al ensayo se le incluye en el estrato que le corresponda en función del sexo y estado civil (ej.: hombre soltero) y dentro de cada estrato se lleva a cabo una aleatorización simple o, más comúnmente, por bloques, para evitar desequilibrios en el tamaño de los grupos de intervención dentro de cada estrato.

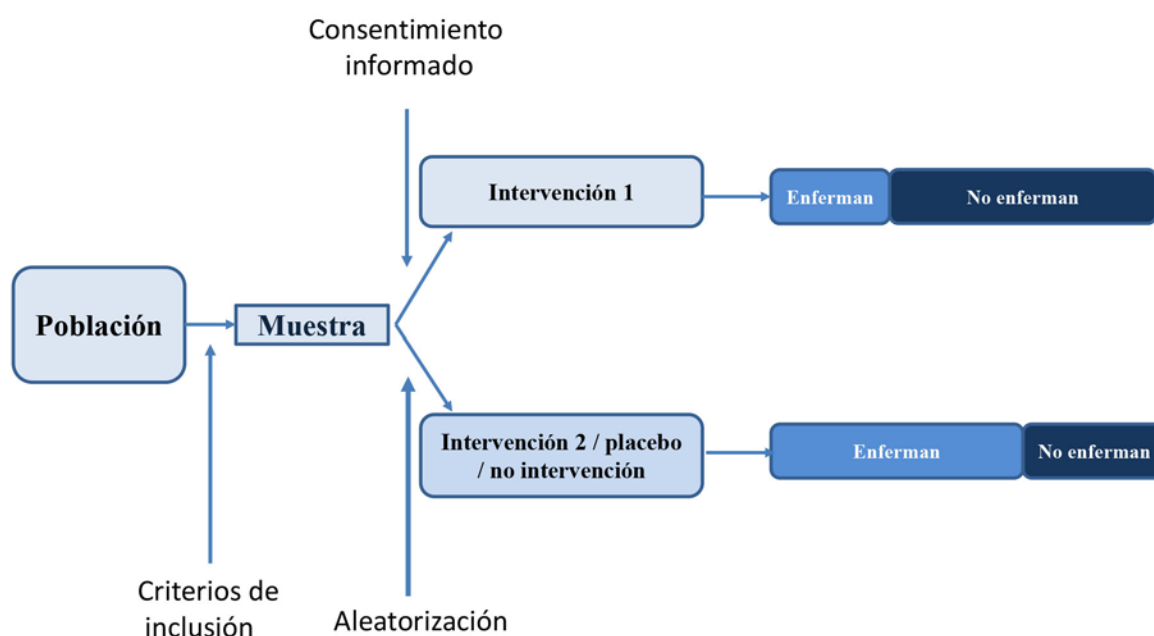
6. Diseño del ensayo

6.1. Ensayo de grupos paralelos

El ensayo de grupos paralelos constituye, por su sencillez, el diseño básico de un estudio de intervención. (6,23) En este diseño, cada participante, una vez reclutado y otorgado el consentimiento informado, es asignado de forma aleatoria a una de las intervenciones y, tras un período de seguimiento, se mide la respuesta (Figura 3). Lo habitual es que haya dos grupos, dada la dificultad de reclutar suficiente número de pacientes por grupo cuando se aumenta el número de intervenciones a comparar. A veces, se incluye un período de prueba previo a

la aleatorización, que tiene por objeto disminuir las tasas de incumplimiento de la intervención y abandono del estudio o estabilizar los valores de las variables de respuesta. Durante el período de prueba, todos los individuos son sometidos a la misma intervención, habitualmente la del grupo control, excluyendo del estudio a los sujetos no cumplidores. Por ej., un ensayo para comparar el efecto de los ácidos grasos monoinsaturados y poliinsaturados sobre el perfil lipídico plasmático incluyó un periodo de prueba en el que todos los sujetos recibieron una dieta rica en ácidos grasos saturados, lo que permitió valorar el grado de adherencia a la dieta y estabilizar los lípidos plasmáticos. (24) A continuación, los sujetos con buena adherencia fueron asignados aleatoriamente a una dieta rica en ácidos grasos mono o poliinsaturados.

Figura 3. Ensayo de grupos paralelos.



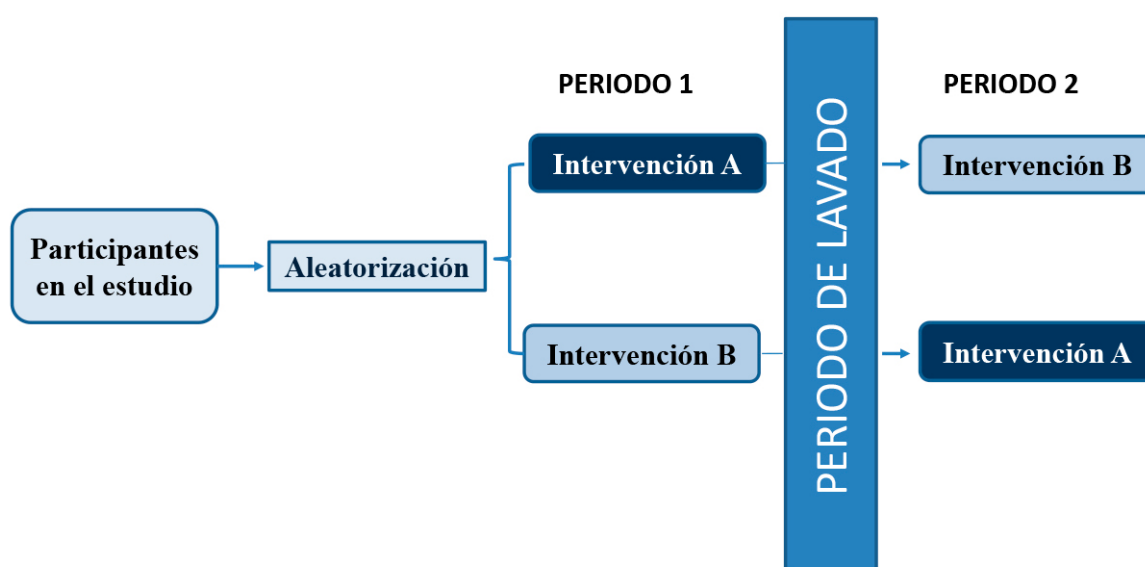
6.2. Ensayo cruzado

El problema principal de los ensayos de grupos paralelos es que se requieren muchos participantes para estimar con precisión los efectos de las intervenciones, comúnmente de pequeña magnitud. Una alternativa consiste en diseñar un ensayo cruzado, en el que cada participante es sometido a todas las intervenciones bajo estudio en periodos de tiempo sucesivos, realizándose después una comparación de intervenciones intra-sujetos. Como cada individuo ejerce como su propio control, se elimina la varianza intraindividual y se precisa una muestra más pequeña que en un ensayo de grupos paralelos para conseguir la misma potencia. Además, los participantes pueden expresar sus preferencias personales por una u otra intervención, en base a su propia experiencia. Como contrapunto, una muestra pequeña puede

condicionar la eficacia de la aleatorización para distribuir las variables pronósticas entre los grupos.

En el diseño cruzado más común, al que nos referiremos aquí, cada individuo es asignado aleatoriamente a una secuencia temporal de dos intervenciones (diseño AB/BA); es decir, lo que se decide al azar no es el grupo de intervención al que será asignado el participante, sino el orden en el que recibirá las intervenciones a comparar (Figura 4). Las variables resultado suelen tener carácter intermedio (no incluyen la curación o la muerte), por lo que este diseño es más útil en evaluaciones iniciales de la eficacia de tratamientos (fase II de ensayos clínicos), siendo de elección en el estudio de bioequivalencia de diferentes formulaciones o dosis del mismo fármaco en voluntarios sanos.

Figura 4. Ensayo cruzado.



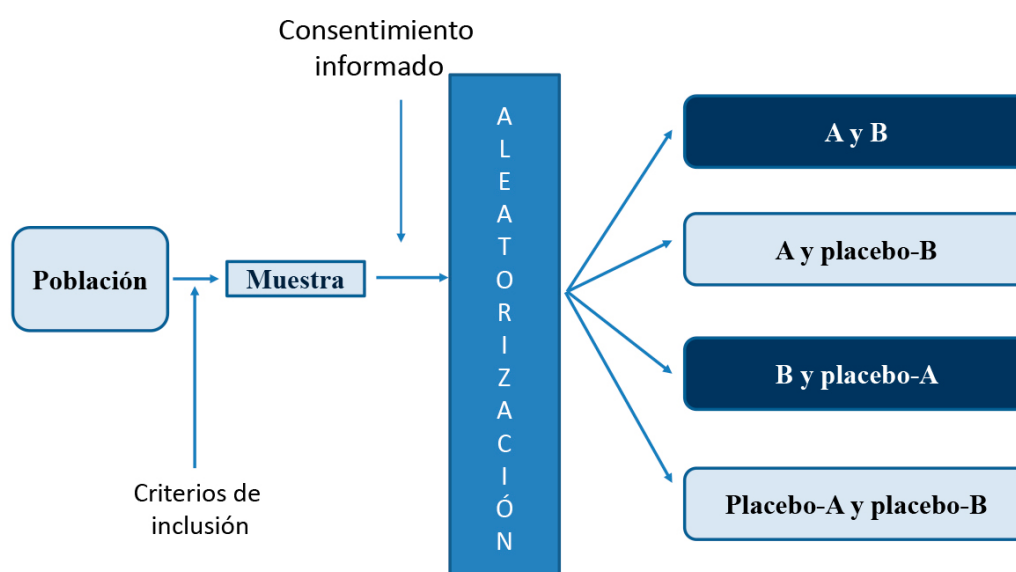
La principal limitación de este diseño es que las condiciones de aplicación son muy restrictivas: especialmente indicado en enfermedades estables con tratamientos no curativos de respuesta rápida y transitoria, requisitos que se dan en algunos procesos crónicos y recidivantes como las enfermedades reumáticas y alérgicas. Además, asume la ausencia de efectos residuales de la intervención. Para garantizar que los efectos del primer tratamiento no persisten durante la administración del segundo (*carry-over effect*) es imprescindible intercalar una fase de lavado (periodo sin tratamiento activo) entre los dos períodos. Otra limitación tiene que ver con los posibles abandonos del estudio durante el segundo período, bien porque el primer tratamiento ha resultado ineficaz o muy eficaz, bien a causa de sus efectos adversos. Estos participantes no deben incluirse en el análisis del ensayo.

Un aspecto a considerar en los ensayos cruzados es el posible efecto período, cuando la condición bajo estudio se modifica con el paso del tiempo o por efecto de variables predictoras inestables, que cambian de un periodo a otro, independientemente de la intervención aplicada. En el diseño AB/BA, con el mismo número de participantes asignados a cada una de las secuencias de intervención, en promedio el efecto período no sesgará el estimador del efecto del tratamiento. Sin embargo, afectará al estimador de la varianza, porque cambia la parte del efecto observado que puede atribuirse al error aleatorio, lo que debe ser tenido en cuenta en los análisis. En un ensayo cruzado hay que usar test estadísticos para datos emparejados, como el t-test o el análisis de la varianza (con efectos participante, período y tratamiento) para medidas repetidas, en el caso de variables de resultado continuas, y el test de Mainland-Gart o la regresión logística condicional, en el caso de variables de resultado dicotómicas. (25)

6.3. Ensayo factorial

Cuando se desea evaluar el efecto de varias intervenciones en un mismo estudio, una opción es diseñar un ensayo de grupos paralelos con un brazo para cada una de las intervenciones y otro brazo para el grupo control: placebo, práctica/tratamiento habitual o no intervención. Por ejemplo, un ensayo de prevención cardiovascular con cuatro brazos para evaluar una intervención dietética (uso exclusivo de aceite de oliva para cocinar y aliñar la comida), un programa de actividad física y una intervención educativa sobre técnicas psicológicas para afrontar el estrés. El número de participantes necesarios no es así tan alto como el que se requeriría al hacer un ensayo distinto para cada intervención.

Figura 5. Ensayo factorial.



El diseño factorial permite evaluar en un estudio varias intervenciones con mecanismos de acción y efectos independientes sin necesidad de aumentar el tamaño muestral requerido para una sola intervención. En un ensayo factorial cada participante es asignado aleatoriamente a una combinación de dos o más intervenciones o de diferentes dosis o niveles de intervención. (6) En el caso más sencillo, con dos intervenciones activas (A y B), el ensayo consta de cuatro brazos: ambas intervenciones, intervención A más placebo, intervención B más placebo y ambos placebos (Figura 5). Suponiendo que el tamaño muestral requerido para evaluar cada intervención por separado fuese de 2000 participantes (1000 en el grupo de intervención y 1000 en el grupo control), el número de participantes necesario para evaluar las dos intervenciones será diferente en función del diseño elegido (Tabla 1):

1. Dos ensayos de grupos paralelos con dos brazos (intervención y placebo) cada uno: 4000 participantes.
2. Un ensayo de grupos paralelos con tres brazos (A, B y placebo): 3000 participantes.
3. Un ensayo factorial, donde tanto la intervención A como la B son administradas a la mitad de los sujetos del estudio: 2000 participantes.

El diseño factorial también sirve para evaluar posibles interacciones entre los efectos de las intervenciones estudiadas, pero entonces son precisos tamaños muestrales más elevados, por lo que pierde su atractivo en términos de eficiencia y la interpretación de los resultados se complica.

Tabla 1: Diseños alternativos para evaluar el efecto de dos intervenciones

	Brazo	Número de participantes	Intervención A	Intervención B
Ensayos de grupos paralelos de 2 brazos				
• Ensayo A	1	1000	A	
	2	1000	A-placebo	
• Ensayo B	1	1000		B
	2	1000		B-placebo
Total		4000		
Ensayo de grupos paralelos de 3 brazos				
	1	1000	A	
	2	1000		B
	3	1000	Placebo	
Total		3000		

Tabla 1: Diseños alternativos para evaluar el efecto de dos intervenciones (continuación).

Ensayo factorial				
	1	500	A	B
	2	500	A	B-placebo
	3	500	A-placebo	B
	4	500	A-placebo	B-placebo
Total		2000		

6.4. Ensayo por conglomerados

Los ensayos por conglomerados son aquellos en los que la unidad de aleatorización es el grupo en lugar del individuo; mientras que la medición del desenlace de interés se sigue realizando en los individuos, lo cual no excluye para que en ocasiones también se midan resultados en variables de grupo. Los grupos pueden ser comunidades enteras de gran tamaño, como barrios o ciudades (ensayos comunitarios), o conglomerados de pequeño tamaño, como familias, aulas, escuelas, profesionales de salud (cupos médicos o de enfermería) o centros de salud. Los ensayos por conglomerados están especialmente indicados cuando existe riesgo de contaminación y cuando se evalúan intervenciones que manipulan el entorno (físico o social) o que han sido diseñadas para ser administradas a grupos.

7. Los participantes

La selección de participantes en un ensayo debe llevarse a cabo de forma cuidadosa. En primer lugar, hay que definir con precisión los sujetos susceptibles de participar, mediante los criterios de inclusión y exclusión, dirigidos a conseguir el número de participantes necesario para detectar el efecto esperado, excluir candidatos en los que es previsible una baja adherencia al protocolo y minimizar el riesgo de efectos adversos e interacciones con otros tratamientos. A veces resulta difícil reclutar suficientes sujetos, sobre todo cuando la magnitud del efecto es pequeña, lo que puede dar lugar a un error de tipo 2 (no detectar una diferencia realmen-

te existente entre las intervenciones). Para contrarrestar este problema se pueden seleccionar sujetos de alto riesgo de desarrollar el desenlace de interés, como en el estudio ATBC, diseñado para evaluar la eficacia de los suplementos dietéticos de α -tocoferol y β -caroteno para prevenir el cáncer de pulmón en varones adultos fumadores. (13) Sin embargo, los resultados en sujetos de alto riesgo no se pueden extrapolar sin más a las poblaciones de moderado o bajo riesgo. Por ejemplo, varios ensayos clínicos han demostrado la eficacia de la aspirina en pacientes con enfermedad cardiovascular estable (prevención secundaria), con una reducción del 21 % de eventos cardiovasculares y del 13 % en la mortalidad por todas las causas. (26) Por cada 1000 pacientes tratados durante 3 años, la aspirina causa 9 eventos hemorrágicos mayores, pero evita 36 eventos cardiovasculares y 16 muertes, con un balance beneficio/riesgo claramente positivo. Sin embargo, las cosas son muy diferentes en prevención primaria. Un metanálisis de ensayos de prevención en personas con diabetes encontró una reducción marginal de eventos cardiovasculares, sin beneficio en términos de mortalidad y con un incremento del 30 % en los eventos hemorrágicos mayores. (27) Además, el ensayo ASPREE en ancianos sanos encontró un aumento de la mortalidad total del 14 % y del riesgo de eventos hemorrágicos mayores del 38 %. (28) En otro metanálisis, que incluyó todos los ensayos de más de 1000 participantes, por cada 265 sujetos tratados durante 10 años se previno un evento cardiovascular y por cada 210 se causó una hemorragia mayor. (29)

7.1. Megaensayos

Al objeto de evitar la incertidumbre generada por la disparidad o imprecisión de los resultados de ensayos de limitado tamaño muestral, Yusuf *et al.* propusieron llevar a cabo ensayos de gran tamaño (megaensayos) para evaluar intervenciones de efecto moderado. (30) Si se pretende reclutar un número muy alto de participantes y que estos se adhieran mayoritariamente al protocolo del estudio, la simplicidad es un principio a tener presente en cada componente del ensayo:

1. Los criterios de inclusión serán amplios y los de exclusión mínimamente restrictivos, solo para sujetos que presenten contraindicaciones con la intervención.
2. La hoja de recogida de datos consignará tan solo la información que se utilizará en el análisis, evitando preguntas abiertas o sin relevancia clara para el objeto de la investigación.
3. Las intervenciones a estudiar, fáciles de aplicar y no invasivas, se ajustarán lo más posible a los procedimientos de la práctica habitual.

4. Para evaluar el efecto de la intervención se usarán, siempre que sea posible, variables que se recogen de forma rutinaria en la práctica habitual.
5. Idealmente, el ensayo debería minimizar el número de visitas que los participantes tienen que realizar a los centros donde se lleva a cabo el seguimiento, que deberá realizarse de forma indirecta, mediante encuestas por teléfono, internet o recogida de información en registros disponibles.

Un ejemplo de este tipo de ensayos es el ISIS-4 (4th *International Study of Infarct Survival*), que reclutó 58 050 pacientes con sospecha de infarto agudo de miocardio. (31) Los participantes fueron aleatorizados mediante un diseño factorial a tres intervenciones: captoril por vía oral frente a placebo, mononitrato de isosorbide por vía oral frente a placebo y sulfato de magnesio por vía intravenosa frente a no tratamiento. Este ensayo clínico estableció la ausencia de un efecto beneficioso de la infusión de magnesio, contradiciendo los resultados de dos metaanálisis de 9 ensayos previos de pequeño tamaño, que encontraron una reducción de la mortalidad en la fase aguda del infarto del 55 %, y del LIMIT-2, un ensayo con 2300 pacientes, donde la mortalidad se redujo un 24 %. (32)

Los megaensayos son los predecesores de los ensayos pragmáticos, que pretenden evaluar la efectividad de las intervenciones en las condiciones de aplicación de la vida real, un objetivo para el que resultan pertinentes las consideraciones previas. Un ensayo será más o menos pragmático en función de los siguientes elementos: criterios de elegibilidad, esfuerzos de reclutamiento, lugar donde se realiza, organización y recursos para llevarlo a cabo, aplicación de la intervención, medidas para garantizar la adherencia, intensidad del seguimiento, relevancia del resultado primario para los participantes y en qué medida se incluyen todos los datos en el análisis. (15) Cuanto más se aproximen estos elementos a las condiciones de la práctica habitual, más pragmático será el ensayo, y viceversa.

8. La intervención

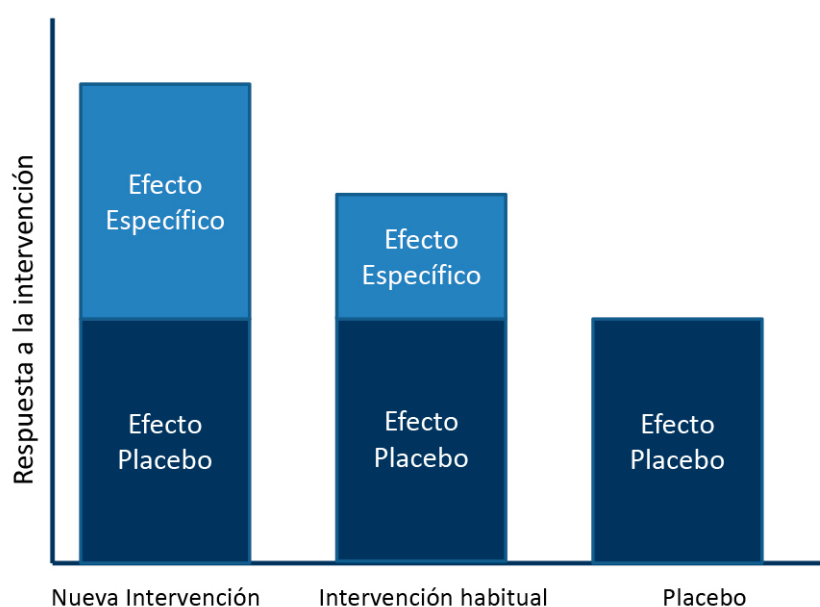
Los ensayos clínicos fueron diseñados inicialmente para evaluar intervenciones farmacológicas. Los tratamientos con fármacos son sencillos de aplicar y estandarizar (dosis, frecuencia, duración y vía de administración). Esto resulta fundamental para reproducir la investigación y transferir sus resultados a la práctica clínica. Además, la administración en el grupo control de un tratamiento inactivo o del tratamiento habitual, en caso de que lo haya, con un aspecto similar al del tratamiento en estudio, permite neutralizar el efecto placebo de forma eficaz.

8.1. El efecto placebo

La administración de una intervención produce un efecto terapéutico sobre la persona que la recibe, denominado efecto placebo, independientemente de la composición o características de la misma. (6) Es decir, incluso la administración de un principio inactivo produce un efecto. El simple hecho de ser observado, al participar en un estudio, produce un efecto en los participantes, denominado efecto *Hawthorne*. (33) A esto hay que añadir el poder de sugestión derivado del hecho de recibir una intervención, aunque ésta sea inactiva, y de la presencia del profesional que la administra.

El efecto global de cualquier intervención tiene dos componentes: el efecto placebo y el efecto específico de la misma (Figura 6). Por ello, es preciso aplicar otra intervención al grupo control (placebo o la práctica habitual), que permita neutralizar el efecto placebo y discernir así el efecto específico de la nueva intervención. Cuando exista una intervención eficaz disponible para la condición bajo estudio, no aplicaremos placebo a los sujetos del grupo control, sino la práctica habitual. El placebo tiene por objeto que la actitud de los individuos ante la investigación sea la misma en los grupos de intervención y control. Asimismo, en algunos tipos de intervenciones, particularmente las farmacológicas, permite enmascarar la intervención, contribuyendo a evitar otro tipo de sesgos en el desarrollo, análisis y evaluación del ensayo, como veremos más adelante.

Figura 6. Efecto placebo.



8.2. Las intervenciones no farmacológicas

El ámbito de la prevención se dirige con frecuencia a la modificación de los estilos de vida (alimentación, actividad física, adicciones, prácticas sexuales) mediante intervenciones no farmacológicas, de carácter complejo, con varios componentes y que con frecuencia son aplicadas por múltiples profesionales y a varios niveles. Esto plantea una serie de problemas metodológicos: (34)

1. Dificultad de describir y estandarizar la intervención, descomponiéndola en sus elementos básicos.
2. Dificultad de adherencia al protocolo por los participantes y los responsables de administrar la intervención.
3. Diferencias en la preparación y experiencia previa de los profesionales que administran la intervención y los centros participantes.
4. Dificultad o imposibilidad de enmascarar la intervención.
5. Demoras entre la asignación de la intervención y su administración, relacionadas con aspectos prácticos, como la formación de los profesionales en la técnica de la intervención o la organización de la agenda de visitas para las sesiones de intervenciones conductuales. Estos retrasos pueden causar contaminación entre grupos, pérdidas de seguimiento y falta de adherencia a la intervención.

La pobre descripción de las intervenciones no farmacológicas dificulta la reproducción de los ensayos y la aplicación de sus resultados en la práctica habitual, lo que ha motivado el desarrollo de una guía para la descripción completa de las intervenciones y su posterior reproducibilidad (*TIDieR–Template for Intervention Description and Replication*), (35) con una lista de comprobación que incluye los siguientes ítems, que se describen con detalle en el anexo 2:

1. Breve denominación.
2. Porqué.
3. Qué (materiales).
4. Qué (procedimientos).
5. Quién la aplica.
6. Cómo se administra.
7. Dónde se aplica.
8. Cuándo y con qué frecuencia e intensidad.
9. Adaptación o personalización.
10. Modificaciones.

Otro aspecto diferencial de los ensayos con intervenciones complejas tiene que ver con la intervención aplicada en el grupo control. Cuando se evalúan vacunas o intervenciones farmacológicas preventivas, el uso del placebo es menos problemático que en los ensayos clínicos, por lo que debería aplicarse salvo que existan razones de peso que lo desaconsejen. Por ejemplo, los ensayos con acupuntura para tratar el dolor tienden a encontrar un beneficio menor cuando se compara con acupuntura simulada que cuando se compara con no tratamiento. (36) Con intervenciones de cambio de comportamiento, donde no es posible usar placebo, el grupo control recibe la intervención habitual, cuando la hay, o nada. Por ejemplo, en un ensayo para evaluar el efecto de una terapia de grupo para dejar de fumar, los participantes en el grupo control pueden recibir el consejo que se da en la práctica clínica habitual. Otra opción es aplicar la misma intervención de baja intensidad, con diferente duración, dosis, frecuencia o medios de administración, como en los ensayos con terapias sustitutivas de nicotina para dejar de fumar que comparan una combinación de terapias (un parche más una forma de acción corta—chicle o pastilla—) con una sola de esas mismas terapias. (37) Con intervenciones conductuales, pueden variar la frecuencia de las sesiones, su duración o los recursos materiales y humanos disponibles para su administración. Por último, en el grupo control se puede aplicar una intervención más sencilla o con diferentes componentes. Por ejemplo, en el ensayo Predimed-plus se compara una intervención intensiva sobre el estilo de vida (dieta mediterránea tradicional hipocalórica, promoción de actividad física y soporte conductual) con la mera educación sobre dieta mediterránea. (38)

Por último, en los ensayos por conglomerados la intervención puede tener varios componentes: unos aplicados a nivel individual, sea sobre el sujeto del ensayo o el profesional encargado de administrar la intervención, y otros a nivel grupal, sea sobre los centros donde se realiza el ensayo o sobre grupos de profesionales o sujetos participantes en el mismo. Con estas intervenciones multifacéticas y complejas resulta más difícil la estandarización, la aplicación uniforme y la consecución de una buena adherencia.

9. Enmascaramiento

El enmascaramiento, o carácter ciego de un ensayo, es el conjunto de medidas o precauciones que se toman con el fin de que las personas implicadas en su desarrollo (sujetos del ensayo y profesionales que asignan o administran la intervención, realizan el seguimiento, miden la respuesta y analizan los resultados)

desconozcan la intervención a la que ha sido asignado cada participante, ya que tal conocimiento podría introducir sesgos en la ejecución, análisis y evaluación del ensayo. En la medida de lo posible, las intervenciones enmascaradas deberán ser idénticas en forma, tamaño, color, sabor, textura, vía y pauta de administración, diferenciándose únicamente por la sustancia activa que contengan.

Para Pocock tres son los agentes relevantes que intervienen en un ensayo: sujeto o paciente, equipo de intervención y evaluador. (6) Decimos que un ensayo es abierto (*open-label*) cuando todos conocen la intervención asignada; simple ciego cuando el sujeto o paciente está enmascarado; y doble ciego cuando también lo están los responsables de administrar la intervención y evaluar la respuesta (se mantiene esta denominación porque en ocasiones son la misma persona, aunque cuando no lo son también se puede hablar de triple ciego). La expresión «evaluación ciega por terceros» se suele reservar para ensayos abiertos en los que están enmascarados los encargados de realizar los análisis estadísticos.

El enmascaramiento del encargado de medir la respuesta (sujeto del ensayo, profesional que administra la intervención o un tercero) es más importante cuando la variable resultado es subjetiva (valoración de la intensidad del dolor o de una imagen diagnóstica), ya que el conocimiento de la intervención asignada puede influir en la percepción de su efecto. Como hemos mencionado previamente, la acupuntura para tratar el dolor muestra un efecto de menor magnitud cuando aplicamos acupuntura simulada en el grupo control, pues el enmascaramiento de los pacientes permite controlar el sesgo por efecto placebo. (36) Además, si el profesional encargado de administrar la intervención en un ensayo es un entusiasta de la eficacia del nuevo tratamiento bajo estudio, podría transmitir su entusiasmo a los pacientes que lo están recibiendo, alterando aún más la valoración de la variable respuesta. Los resultados más objetivos, como la mortalidad, son relativamente insensibles a este tipo de sesgo.

Lamentablemente, con algunas intervenciones, como las dietéticas o quirúrgicas, entre otras, no siempre es posible aplicar el enmascaramiento. En otras ocasiones, los tratamientos tienen efectos secundarios bien conocidos que permiten la identificación de los mismos, como en el caso de la rifampicina, que tiñe de rojo la orina. Además, nunca deben enmascararse los tratamientos si con ello sometemos a algún tipo de riesgo a los sujetos del ensayo. Por ello, independientemente de la nomenclatura que usemos (simple, doble, triple ciego, etc.), para la que no existe consenso, lo importante es justificar y describir el tipo de enmascaramiento que se ha realizado o las razones que han impedido su aplicación, las posibles violaciones del mismo y otras circunstancias que puedan haber puesto en riesgo su eficacia.

10. Análisis de un ensayo

El protocolo de un ensayo debe presentar un plan de análisis detallado, distinguiendo claramente entre la variable de respuesta principal, que determinará el tamaño muestral y la posible eficacia de la intervención, y las secundarias. (23) En un ensayo sobre los efectos de la dieta mediterránea, el resultado de mayor interés es la incidencia o mortalidad debida a enfermedades crónicas no transmisibles relacionadas con la alimentación, como las vasculares o el cáncer. Sin embargo, para detectar un número suficiente de casos o defunciones sería necesaria una muestra muy numerosa de sujetos de alto riesgo, dispuestos a someterse a seguimiento durante un largo período de tiempo, condiciones que pueden limitar la factibilidad del estudio. A corto o medio plazo, pueden utilizarse variables intermedias de respuesta, como la presión arterial, el índice de masa corporal o el nivel de oxidación de determinados marcadores biológicos.

10.1. Análisis por intención de tratar, por protocolo o por tratamiento

El análisis principal de un ensayo se basa en el principio de intención de tratar, que incluye a cada participante en su grupo de aleatorización (por asignación), incluso cuando una parte de los mismos no se hayan adherido al protocolo, por incumplimiento de la intervención, administración de una intervención distinta a la asignada, pérdida del seguimiento o retirada del estudio. (6) El análisis por asignación presenta dos ventajas. Por un lado, preserva lo conseguido con la aleatorización, que los grupos sean comparables, el fundamento para obtener conclusiones válidas sobre el efecto de una intervención. Por otro lado, es un método pragmático, que pretende analizar el efecto de la intervención de forma similar a lo que es previsible que ocurra en la práctica real (mejor aproximación a la efectividad), en contraste con el análisis por protocolo, que solo incluye a los participantes que se han adherido a la intervención asignada. El análisis por asignación es conservador, por lo que cabe la posibilidad de obtener resultados nulos en presencia de un efecto real de la intervención, sobre todo cuando este es de pequeña magnitud. Este sesgo hacia el valor nulo es debido a la falta de adherencia a la intervención asignada, ya que los participantes asignados a la intervención que no se adhieran a la misma presentarán una frecuencia del desenlace del estudio similar a los del grupo control y los participantes del grupo control que reciban, de forma parcial o completa, la intervención asignada al grupo de intervención, presentarán una frecuencia del desenlace del estudio similar a los de este grupo. Por tanto, si encontramos un efecto beneficioso de la intervención con un análisis por asignación

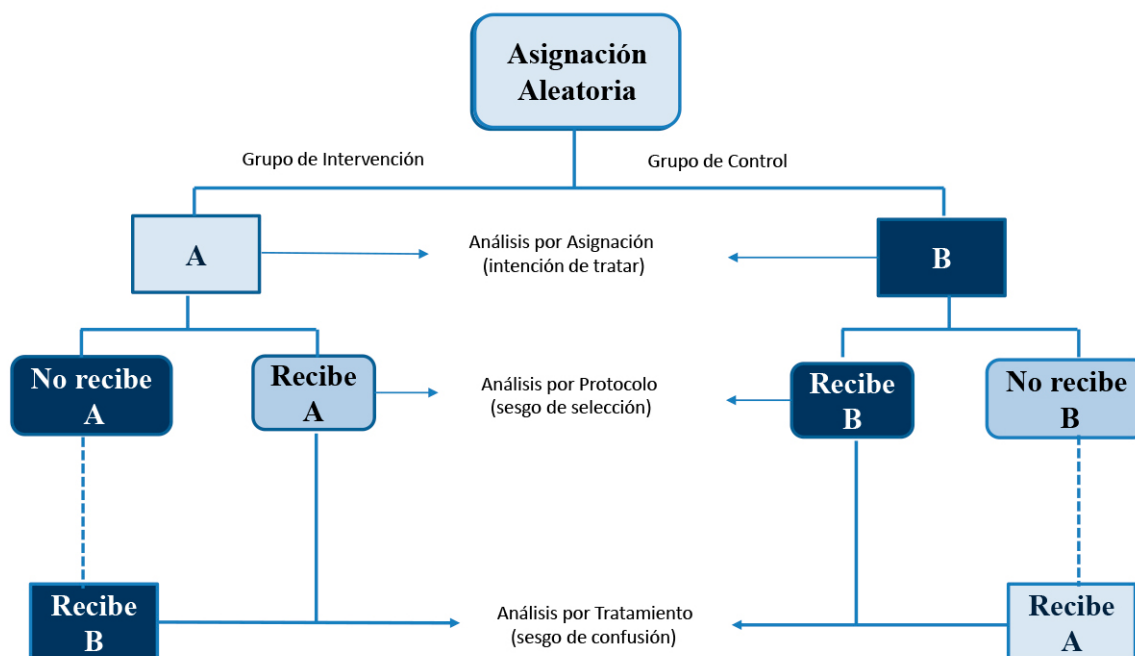
en presencia de contaminación, eso significa que el efecto real de la intervención será aún mayor. En los ensayos de no inferioridad, con dos intervenciones activas, este sesgo puede dar como resultado que la nueva intervención no es inferior a la intervención habitual cuando en realidad sí lo es, si el tipo y la magnitud de la no adherencia es similar en ambos grupos. (39) Y, de forma general, en los ensayos con dos intervenciones activas, sean estos de no inferioridad o de superioridad, este sesgo puede darse también en dirección opuesta al valor nulo, dependiendo de la magnitud y las razones de la no adherencia en cada grupo. (39)

El efecto de la asignación, que obtenemos con el análisis por intención de tratar, depende pues del patrón de adherencia a la intervención del ensayo de los participantes en el mismo. Por tanto, si analizamos la misma intervención en dos ensayos con patrones de adherencia distintos, el efecto resultante también diferirá. Si la adherencia al protocolo fuese perfecta, de modo que todos los participantes recibieran la intervención asignada en las condiciones estipuladas en el protocolo y no se produjese ninguna retirada del estudio, entonces el análisis por intención de tratar nos proporcionaría el efecto real de la intervención, ya que este coincidiría con el efecto de la asignación. El efecto de la asignación tiene interés para ciudadanos, evaluadores, gestores y políticos, puesto que representa una mejor aproximación a la efectividad, lo cual es muy relevante para decidir si se autoriza o financia una intervención. No obstante, el efecto de la asignación no es equivalente a la efectividad, pues ni el perfil de los participantes ni las condiciones en que se realiza un ensayo (monitorización más intensa y enmascaramiento de la intervención) coinciden con las de la aplicación posterior de la intervención en la práctica real. Y el efecto de la intervención tiene interés para los investigadores, que pueden diseñar nuevos estudios en base a esa información, y para las personas susceptibles de someterse a la misma en el futuro, especialmente para las más cumplidoras, pues informa de su eficacia cuando se siguen las indicaciones del protocolo del ensayo.

Existen dos estrategias para estimar el efecto de la intervención en un ensayo: el análisis por tratamiento y el análisis por protocolo. (39) En el análisis por tratamiento, cada participante es clasificado de acuerdo con la intervención finalmente recibida, independientemente de la asignación inicial (sesgo de confusión), y en el análisis por protocolo solo se incluyen aquellos participantes que han recibido la intervención asignada en las condiciones del protocolo (sesgo de selección), dejando fuera a los incumplidores (Figura 7). En ambos casos se rompe el principio de aleatorización, de modo que ya no existe garantía de que los grupos sigan siendo comparables, pues las razones que han llevado a los participantes a adherirse al protocolo o a recibir una u otra intervención suelen estar relacionadas con el desenlace del estudio y los sesgos correspondientes, de selección y confusión, pueden ir en cualquier dirección. Por tanto, a todos los efectos, la situación es la misma que si los datos procedieran de un

estudio observacional, lo que obliga a controlar la posible confusión provocada por los factores asociados con la decisión de adherirse al protocolo o recibir una u otra intervención, mediante el correspondiente análisis ajustado.

Figura 7. Análisis de los ensayos.



10.2. Medición del efecto

Otro aspecto clave del análisis de un ensayo es la elección del indicador para medir el efecto de la intervención. Además de las medidas relativas, como la reducción del riesgo relativo (RRR), es importante proporcionar medidas absolutas, como la reducción absoluta del riesgo (RAR) o diferencia en el riesgo de contraer la enfermedad o condición en estudio entre el grupo control (I_c) y el de intervención (I_i), y el número de pacientes que hay que tratar para evitar un caso de la enfermedad o un evento de la condición en estudio (NNT, por sus siglas en inglés: *Number Needed to Treat*) o para producir un efecto adverso (NNH por sus siglas en inglés: *Number Needed to Harm*), que se calcula como el inverso de la RAR. Las medidas absolutas, y el NNT en particular, son muy útiles para valorar el balance coste-beneficio y beneficio-riesgo de nuevas intervenciones. Las fórmulas para el cálculo de cada uno de estos indicadores son las siguientes:

- *Reducción Absoluta del Riesgo (RAR) o diferencia de riesgos*

$$RAR = I_c - I_i$$
- *Reducción del Riesgo Relativo (RRR) o proporción prevenida en expuestos*

$$RRR = (I_c - I_i) / I_c = 1 - RR; \text{ donde } RR \text{ es la razón de riesgos}$$

La RRR se expresa habitualmente como porcentaje: $(1 - RR) \times 100$

- *Número de pacientes a tratar para evitar un caso o evento (NNT)*

$$NNT = 1 / (I_c - I_i) = 1 / RAR$$

El NNH se calcula como el inverso del aumento absoluto del riesgo (AAR)

- *Aumento Absoluto del Riesgo (AAR)*

$$AAR = I_i - I_c$$

- *Número de pacientes a tratar para producir un efecto adverso (NNH)*

$$NNH = 1 / (I_i - I_c) = 1 / AAR$$

La RRR puede producir una impresión magnificada del efecto del tratamiento, por lo que se recomienda presentar esta medida acompañada de alguna magnitud de la reducción del riesgo en términos absolutos, particularmente el NNT. Para ilustrar esto, a partir de una revisión de ensayos con tratamientos farmacológicos para la hipertensión arterial, se analizaron por separado los efectos en sujetos con hipertensión ligera y moderada. (40) La reducción del riesgo relativo de padecer un accidente cerebrovascular (ACV) en 5 años, fue igual en los dos grupos, del 40 %, correspondiente a un RR de 0,6 (Tabla 2). Sin embargo, la reducción absoluta del riesgo fue mucho menor en los sujetos con hipertensión ligera, dado que estos tenían una incidencia acumulada (denominada tasa en la tabla) de ACV en ausencia de tratamiento mucho más baja que los del grupo de hipertensión moderada. Traducido en términos de NNT, mientras es preciso tratar a 167 sujetos con hipertensión ligera para evitar un ACV en 5 años, basta con tratar a 13 sujetos con hipertensión moderada para obtener el mismo resultado.

Tabla 2. Cálculo de la reducción del riesgo relativo (RRR), la reducción absoluta del riesgo (RAR) y el número de sujetos a tratar (NNT) para evitar un caso de accidente cerebrovascular (ACV) en 5 años

	ACV en 5 años		RRR	RAR	NNT
Hipertensión	Grupo control	Grupo tratamiento	Pc – Pt) / Pc	Pc – Pt	1 / (Pc – Pt)
Moderada					
Tasa de ACV (P)	0,20	0,12	0,40	0,08	13
N	16 778	16 898			
Ligera					
Tasa de ACV (P)	0,015	0,009	0,40	0,006	167
N	15 165	15 238			

Fuente: Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310:452- (4)

10.3. Análisis en subgrupos

Si los sujetos participantes en un ensayo responden de forma diferente a la intervención en función de determinadas características (por ej. sexo, edad, genotipo, nivel educativo o socioeconómico), el efecto global de la misma pierde parte de su interés y puede inducir a error. Para detectar esta modificación del efecto es necesario realizar análisis en subgrupos definidos por las características que podrían determinar una respuesta diferencial a la intervención, pero habitualmente el tamaño muestral de los ensayos se estima en función del efecto global esperado, por lo que suelen carecer de potencia para detectar posibles interacciones estadísticas.

En muchas ocasiones, cuando se ha detectado la presencia de modificación del efecto de una intervención en un ensayo, a la postre se ha demostrado que se trataba de un error. (41) En la mayor parte de los casos, se trataba de un error aleatorio, pero también se han publicado resultados falsos por sesgos debidos a la selección del resultado publicado (ver el siguiente apartado), escasez de conocimientos y datos previos sugestivos de modificación del efecto o uso de métodos estadísticos inapropiados.

Para abordar este problema, se han definido varios criterios de credibilidad de la presencia de modificación del efecto. Un grupo de investigadores liderados por Stefan Schandelmaier han revisado la literatura científica para identificar los criterios publicados y, mediante un procedimiento de consenso con expertos, han desarrollado un instrumento para valorar la credibilidad de los análisis de modificación del efecto en los ensayos clínicos, (41) que incluye 5 preguntas:

1. ¿Se planteó correctamente la hipótesis de la modificación del efecto *a priori*?
2. ¿Existían pruebas científicas previas que respaldaran la modificación del efecto?
3. ¿El test estadístico de interacción sugiere que el azar es improbable como explicación de la aparente modificación del efecto?
4. ¿Analizaron los autores un número pequeño de modificadores del efecto o consideraron el número de test realizados en su análisis estadístico?
5. ¿Si el modificador del efecto es una variable continua, se evitaron los puntos de corte arbitrarios?
6. ¿Hay alguna consideración adicional que puede aumentar o reducir la credibilidad?

Cada pregunta tiene 4 opciones de respuesta, de menor a mayor grado de apoyo a la credibilidad. El instrumento concluye con una evaluación general de la credibilidad en una escala analógica visual dividida en 4 áreas (credibilidad muy baja, credibilidad baja, credibilidad moderada y alta credibilidad). Las 4 áreas co-

responden aproximadamente a probabilidades de menos del 25 %, 25 %-50 %, 51 %-75 % y mayor que 75 %, respectivamente, de que la modificación del efecto realmente exista.

11. Sesgos en los ensayos

La aleatorización y el enmascaramiento son los principales mecanismos para asegurar la intercambiabilidad de los grupos comparados en un ensayo. No obstante, la aleatorización puede fallar por diversos motivos y el enmascaramiento no siempre es factible. Además, es posible que algunos participantes no se adhieran al protocolo del estudio o se pierdan durante el seguimiento del mismo. Todas estas circunstancias pueden dar lugar a sesgos que comprometen la validez interna de los ensayos. La colaboración Cochrane clasifica los posibles sesgos de los ensayos, en función de su mecanismo de producción, en cinco dominios: 1) aleatorización, 2) desviación de la intervención asignada, 3) información sobre el desenlace de interés, 4) medición del resultado y 5) selección del resultado publicado. (42)

11.1. Sesgos surgidos durante el proceso de aleatorización

Ya se ha mencionado que si el número de participantes por grupo es bajo (< 100) la aleatorización no garantiza la distribución homogénea de los factores pronósticos en los grupos comparados. La forma de prevenir sesgos durante este proceso es utilizar un tamaño de muestra suficiente (al menos 300 sujetos por grupo) y un sistema de asignación aleatoria remoto y centralizado. Existirá riesgo de sesgo de selección cuando el investigador encargado del reclutamiento pueda predecir la secuencia de asignación, conociendo así con antelación la intervención a la que será asignado el próximo sujeto. Esto puede darse cuando se utilizan métodos de aleatorización sistemáticos (no aleatorios, ya que no están basados en propiedades matemáticas conocidas ni son reproducibles), con asignación según el orden de incorporación al estudio, la fecha de nacimiento, la primera letra del apellido o el día de la semana, y cuando se utiliza la aleatorización por bloques de tamaño prefijado. En esta modalidad, los últimos sujetos de cada bloque no se asignan al azar, sino al grupo que ha recibido menos asignaciones dentro de cada bloque. Si el tamaño del bloque se conoce o se desvela (observando las sucesivas asignaciones de los participantes), se puede adivinar a qué grupo serán asignados los últimos

sujetos de cada bloque. La presencia de desequilibrios importantes en el tamaño de los grupos comparados o en la distribución de los factores pronósticos (tabla 1 del ensayo) es sugestiva de fallos en el procedimiento de aleatorización: mayor número de diferencias estadísticamente significativas en las características basales de los grupos de comparación que las esperadas por azar (5 %), valores P muy pequeños ($<0,001$) o diferencias de magnitud suficiente para introducir confusión capaz de sesgar el estimador del efecto. La similitud excesiva de los grupos, más allá de lo esperable por azar, también es sugestiva de fallos en la aleatorización.

11.2. Sesgos debidos a desviaciones de la intervención asignada

Se entiende por desviación de la intervención asignada a cualquiera de las siguientes:

1. Administración de intervenciones adicionales no previstas en el protocolo. No se consideran desviaciones los cambios en las intervenciones contemplados en el protocolo, como la cesación o modificación de un tratamiento por efectos secundarios, o la aplicación de un aumento de dosis o una segunda línea de tratamiento, por progresión de la enfermedad.
2. Los profesionales no administran la intervención según la pauta programada.
3. Falta de adherencia a la intervención por los participantes: adherencia incompleta (uso intermitente, dosis más baja de la recomendada o rechazo de algún componente en intervenciones complejas), cesación completa de la intervención, contaminación (administración de la intervención asignada al grupo de comparación) o cambio a otra intervención activa.

Los ensayos abiertos o aquellos en los que falla el enmascaramiento son especialmente susceptibles a estas desviaciones, por lo que, en tales casos, conviene ser rigurosos en la aplicación de la intervención y el seguimiento de los pacientes para minimizar el riesgo de sesgos.

Cualquier desviación de la intervención asignada puede producir un sesgo cuando se analiza el efecto de la intervención (análisis por protocolo o por tratamiento). Sin embargo, cuando se analiza el efecto de la asignación (análisis por intención de tratar), estas desviaciones pueden dar lugar a sesgos únicamente cuando son fruto del contexto del ensayo, relacionadas con las actividades de reclutamiento y seguimiento de los participantes cuando las intervenciones son abiertas o hay fallos en el enmascaramiento. Así, el abandono voluntario de la intervención asignada por un participante no produce un sesgo en la estimación del efecto de la asignación a la intervención. Se asume que, en presencia de enmascaramiento, las desviaciones de la intervención asignada no son debidas al contexto del ensayo.

11.3. Sesgos debidos a la ausencia de información sobre el desenlace de interés (datos ausentes)

La ausencia de datos sobre el desenlace de interés del estudio puede deberse a una o varias de las siguientes razones:

1. Pérdidas de seguimiento o retiradas del estudio.
2. No asistencia a una visita en la que debía medirse la variable resultado.
3. Fallecimiento por otras causas.

Cuando esta falta de información esté relacionada con el desenlace de interés se producirá un sesgo; es decir, cuando la probabilidad de desarrollar el desenlace de interés sea diferente entre aquellos para los que se dispone o no de información sobre el mismo (datos ausentes no aleatorios). La presencia de un porcentaje diferente de datos ausentes en los grupos comparados es sugestiva de la presencia de sesgo. En los estudios de supervivencia, la inclusión en los análisis del tiempo de seguimiento observado en los participantes truncados no excluye el riesgo de sesgo si las pérdidas al seguimiento o retiradas del estudio están relacionadas con el desenlace de interés.

11.4. Sesgos debidos a errores en la medición del resultado

Los errores en la medición del resultado, de mala clasificación (variables dicotómicas o categóricas) o de detección de eventos (infra o sobrediagnóstico) son especialmente problemáticos cuando son diferenciales en función del grupo de asignación. Los errores de medición no diferenciales no suelen causar sesgos cuando el resultado se mide con variables continuas (diferencias de medias) y suelen atenuar el estimador del efecto (sesgo hacia el valor nulo) cuando este se mide con razones de *odds* (*odds ratios*-OR), razones de riesgos (*risk ratios*-RR) o razones de riesgos instantáneos (*hazard ratios*-HR). Existen situaciones en las que un error de medición no diferencial puede sesgar el estimador del efecto en otra dirección, (43) pero es poco probable que esto ocurra en un estudio de intervención, por lo que centraremos nuestra atención en los errores de medición diferenciales.

Los estudios de intervención abiertos son especialmente sensibles a este tipo de sesgos, que se previenen con el enmascaramiento de los encargados de medir el resultado. La ausencia de enmascaramiento se asocia, en promedio, con una sobreestimación de las OR del 34 %; (44) aunque el riesgo de sesgo es mayor con variables sujetas al juicio del observador, como escalas de dolor o síntomas depresivos, que con otras más objetivas, como la mortalidad. El riesgo de sesgo también es mayor cuando la medición del resultado depende de los participantes en el ensayo o de los profesionales que administran la intervención que cuando es llevada a cabo por

otro investigador. Por último, la aplicación de una intervención activa en el grupo control reduce el riesgo de este tipo de sesgo, ya que los participantes no tendrán una predisposición a creer que una intervención es mejor que la otra, cosa que sí puede ocurrir cuando el grupo control no recibe ninguna intervención o esta se limita al tratamiento habitual.

11.5. Sesgos debidos a la selección del resultado publicado

En este apartado distinguimos, por un lado, los sesgos debidos a la no publicación de parte de los resultados de un estudio y, por otro lado, los sesgos debidos a la selección del método de análisis y presentación del resultado. Para ilustrarlo con un ejemplo, imaginemos un estudio sobre un nuevo tratamiento para la ansiedad y la depresión. (42) Si los investigadores solo publicasen los resultados de la depresión, estaríamos ante un sesgo por no publicación de parte de los resultados, los relativos a la ansiedad. Pero los resultados publicados también pueden estar sesgados si el método de análisis y presentación de los mismos no es el apropiado. Supongamos que la depresión se ha medido con dos escalas diferentes, pero presentamos únicamente los datos de una de las escalas, la que arrojó en un momento dado un resultado más favorable al nuevo tratamiento, o los de solo una parte del seguimiento, cuando se observó un estimador del efecto más grande, pese a haber medido la variable resultado en diferentes momentos (a los 3 meses, a los 6 meses, al año, etc.), o los ajustados por un subgrupo de variables determinado, el único modelo con el que se obtuvo un resultado estadísticamente significativo. El registro y publicación de los protocolos de los ensayos, con el plan de análisis completo y los resultados previstos, es esencial para prevenir este tipo de sesgos o, en caso de producirse, poder detectarlos. Los protocolos también tienen que describir como se recogerán y analizarán los posibles efectos adversos de la intervención, tanto los previstos como los imprevistos. El estudio COMPARE cotejó los resultados publicados en una muestra de 67 ensayos clínicos de 5 revistas de alto impacto con las variables de respuesta primarias y secundarias especificadas en los protocolos. (45) En promedio, cada ensayo publicó el 58,2 % de los resultados previstos y una media de 5,3 resultados no contemplados en los protocolos.

11.6. Otros sesgos

Otras dos posibles fuentes de sesgo que hay que tener en cuenta en cualquier ensayo son las siguientes:

1. Sesgo por conflictos de intereses de los autores, habitualmente derivados de sus vínculos financieros con compañías interesadas en los resultados del

estudio. Varias revisiones sistemáticas han puesto de manifiesto que los estudios financiados por las industrias farmacéutica y de la alimentación obtienen resultados más favorables a sus productos que los estudios independientes. (46,47)

2. Sesgo por detención precoz del ensayo cuando los análisis intermedios detectaron un beneficio de la intervención. Una revisión sistemática de ensayos detenidos precozmente encontró que la mayoría estaban financiados por la industria farmacéutica, no declaraban información relevante sobre la decisión de detener el ensayo y hallaban efectos muy grandes, poco plausibles, particularmente cuando el número de eventos era pequeño. (48) Esto es similar a lo observado con ensayos de pequeño tamaño muestral que encuentran efectos grandes que no son confirmados por estudios posteriores, y que cabe atribuir al azar. (49) Los ensayos solo deberían detenerse precozmente cuando existen pruebas abrumadoras sobre el beneficio de la intervención (un valor *p* altamente significativo) basadas en un número suficientemente grande de eventos. (50)

12. La declaración CONSORT

La importancia de una adecuada presentación de los resultados de los ensayos condujo a la elaboración y publicación en 1996 de la declaración CONSORT (*CONsolidated Standards Of Reporting Trials*) y a su posterior revisión y actualización en 2001 y 2010. (51) La declaración recoge un conjunto mínimo de recomendaciones para mejorar la calidad de la presentación de los resultados de ensayos aleatorizados de grupos paralelos elaboradas por investigadores, estadísticos, epidemiólogos y editores biomédicos con el apoyo de numerosas revistas biomédicas y grupos editoriales.

La declaración CONSORT comprende un diagrama de flujo de los participantes del ensayo y una lista de comprobación de 25 ítems con todos los aspectos relevantes para saber cómo fue diseñado, analizado e interpretado un ensayo (anexo 3). La declaración se acompaña de un documento con una explicación detallada de cada ítem y de cómo deben aplicarse en el informe de un ensayo. (52) La declaración original está pensada para el informe de un ensayo estándar con un diseño de dos grupos paralelos. Para facilitar la presentación de ensayos clínicos con metodologías diferentes, se han elaborado una serie de extensiones en función de los siguientes criterios:

1. Tipo de diseño: extensiones para ensayos por conglomerados, pragmáticos, de grupos paralelos de más de dos brazos y de no inferioridad o equivalencia, entre otros.
2. Tipo de intervención: extensiones para intervenciones no farmacológicas, acupuntura e intervenciones sociales y psicológicas, entre otras.
3. Tipo de datos: extensiones para ensayos cruzados, para daños y para resúmenes, entre otras.

Toda la información sobre la declaración CONSORT y sus extensiones, junto con ejemplos y recursos útiles puede consultarse en <https://www.equator-network.org/reporting-guidelines/consort/>

13. Conclusión

El patrón de referencia para evaluar la eficacia y seguridad de una intervención es el ensayo controlado y aleatorizado. No obstante, no existe ningún estudio perfecto, y los ensayos, aunque en menor medida que otros tipos de estudios epidemiológicos, también pueden verse afectados por varios tipos de sesgos. Además, los ensayos tienen sus propias limitaciones, como el déficit de validez externa, su complejidad logística, que limita su uso en situaciones de urgencia, como los brotes de enfermedades transmisibles, y la dificultad de reclutar muestras de tamaño grande durante periodos de tiempo suficientes, por lo que son poco apropiados para el estudio de enfermedades raras, la valoración del efecto a largo plazo de algunas intervenciones, como las dietéticas, y la identificación de efectos adversos raros, pero graves. (53) Por último, existen dos condiciones en las que no es factible la realización de un estudio de intervención o solo puede hacerse bajo criterios muy restrictivos: 1) por razones éticas, cuando estudiamos el efecto de posibles factores de riesgo para la salud, como las bebidas alcohólicas o azucaradas; 2) por razones de factibilidad, relacionadas con cuestiones de tipo ético y práctico, cuando se trata de evaluar el efecto de políticas de salud pública, como las medidas legislativas para regular la oferta y la demanda de productos no saludables (leyes antitabaco, restricciones a la venta y consumo de alcohol, impuesto a las bebidas azucaradas, criterios nutricionales y alimentarios para menús escolares, campañas de marketing social, etc.). Por tanto, la toma de decisiones en salud pública debe guiarse por los mejores métodos científicos disponibles, mediante la realización de estudios de intervención siempre que sea factible, junto con otro tipo de estudios epidemioló-

gicos. La combinación de todas las fuentes de información disponibles, teniendo en cuenta las fortalezas y limitaciones de cada tipo de estudio, constituye la mejor garantía para establecer recomendaciones de salud pública sensatas. Tan poco prudente resulta establecer una recomendación sin suficientes pruebas científicas que la avalen como dejar la decisión para más adelante porque los datos disponibles no son perfectos, cuando la alternativa de no actuar o de continuar con la práctica habitual es peor, en términos de los balances de riesgo-beneficio y coste-efectividad, en base a las mejores pruebas científicas disponibles en ese momento.

14. Referencias

1. Diccionario de la lengua española. Real Academia Española. Disponible en: <https://dle.rae.es/>
2. Asimov I. El cometa Halley. Editorial Círculo de Lectores. Madrid, 1986.
3. Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
4. Hill AB. The Clinical Trial. N Engl J Med. 1952;247:113-9.
5. Bull JP. The historical development of clinical therapeutic trials. J Chronic Dis. 1959;10:218-48.
6. Pocock SJ. Clinical trials. A practical approach. Chichester, GB: John Wiley & Sons Ltd, 2013.
7. Daniels M, Hill AB. Chemotherapy of pulmonary tuberculosis in young adults; an analysis of the combined results of three Medical Research Council trials. Br Med J. 1952;1:1162-8.
8. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. Lancet. 1988;2:349-60.
9. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. Ann Intern Med. 2000;133:455-63.
10. Ford I, Norrie J. Pragmatic Trials. N Engl J Med. 2016;375:454-63.

11. Francis TJ, Korn R, Voight RB, *et al.* An evaluation of the 1954 poliomyelitis vaccine trials. *Am J Public Health Nations Health.* 1955;45:1-63.
12. Franceschi S, Plummer M. Intervention trials. En: *Handbook of epidemiology*, 2nd edition. Ahrens W, Pigeot I (eds.). New York: Springer 2014:365-88.
13. Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med.* 1994;330:1029-35.
14. Peto R, Doll R, Buckley JD, *et al.* Can dietary beta-carotene materially reduce human cancer rates? *Nature.* 1981;290:201-8.
15. Loudon K, Treweek S, Sullivan F, *et al.* The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ.* 2015;350:h2147.
16. Tetzlaff JM, Chan A-W, Kitchen J, Sampson M, Tricco A, Moher D. Guidelines for randomized clinical trial protocol content: a systematic review. *Syst Rev.* 2012;1:43.
17. Chan AW, Tetzlaff JM, Altman DG, *et al.* SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann Intern Med.* 2013;158:200-7.
18. Chan AW, Tetzlaff JM, Gøtzsche PC, *et al.* SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ.* 2013;346:e7586.
19. Real Decreto 1090/2015, de 4 de diciembre, por el que se regulan los ensayos clínicos con medicamentos, los Comités de Ética de la Investigación con medicamentos y el Registro Español de Estudios Clínicos. BOE núm. 307, de 24 de diciembre de 2015, páginas 121923 a 121964.
20. Galende I, Sacristán JA, Soto J. Cómo mejorar la calidad de los ensayos clínicos. *Med Clin (Barc).* 1983;80:768-71.
21. Farrin A, Russell I, Torgerson D, Underwood M; UK BEAM Trial Team. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clin Trials.* 2005;2:119-24.
22. Martínez-González MÁ, Toledo JB y López-Fidalgo J. Intervalos de confianza y contraste de hipótesis. En: Martínez-González MÁ, Sánchez-Villegas A, Toledo E y Faulin J (eds.). *Bioestadística amigable.* Barcelona: Elsevier, 2014:100-46.
23. Guallar E, Royo-Bordonada MA. Directrices del “Committee proprietary medicinal products”: Estadística. En Ergón ed. *II Reunión de Actualización de ensayos clínicos con medicamentos.* Madrid, 1996.
24. Mensink RP, Katan MB. Effect of a diet enriched with monounsaturated or polyunsaturated fatty acids on levels of low-density and high-density

- ty lipoprotein cholesterol in healthy women and men. *N Engl J Med*. 1989;321:436-41.
25. Dwan K, Li T, Altman DG, Elbourne D. CONSORT 2010 statement: extension to randomised crossover trials. *BMJ*. 2019;366:l4378.
 26. Berger JS, Brown DL, Becker RC. Low-dose aspirin in patients with stable cardiovascular disease: a meta-analysis. *Am J Med*. 2008;121:43-9.
 27. Caldeira D, Alves M, David C, Costa J, Ferreira JJ, Pinto FJ. Aspirin in the primary prevention of cardiovascular disease on diabetic patients: Systematic review and meta-analysis. *Prim Care Diabetes*. 2020;14:213-21.
 28. McNeil JJ, Nelson MR, Woods RL, Lockery JE, Wolfe R, Reid CM, *et al*. Effect of Aspirin on All-Cause Mortality in the Healthy Elderly. *N Engl J Med*. 2018;379:1519-28.
 29. Zheng SL, Roddick AJ. Association of Aspirin Use for Primary Prevention With Cardiovascular Events and Bleeding Events: A Systematic Review and Metaanalysis. *JAMA*. 2019;32:277-87.
 30. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med*. 1984;3:409-22.
 31. Four International Study of Infarct Survival Collaborative Group. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. *Lancet*. 1995;345:669-85.
 32. Visser PJ, Bredero AC, Hoekstra JB. Magnesium therapy in acute myocardial infarction. *Neth J Med*. 1995;46:156-65.
 33. Mayo E. The human problems of an industrial civilization. Volume 3. 2nd edition. New York, MacMillan; 1993:53-73.
 34. Boutron I, Altman DG, Moher D1 Schulz KF1, Ravaud P; CONSORT NPT Group. CONSORT Statement for Randomized Trials of Nonpharmacologic Treatments: A 2017 Update and a CONSORT Extension for Nonpharmacologic Trial Abstracts. *Ann Intern Med*. 2017;167:40-7.
 35. Hoffmann TC, Glasziou PP, Boutron I, *et al*. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014;348:g1687.
 36. Madsen MV, Gøtzsche PC, Hróbjartsson A. Acupuncture treatment for pain: systematic review of randomised clinical trials with acupuncture, placebo acupuncture, and no acupuncture groups. *BMJ*. 2009;338:a3115.

37. Lindson N, Chepkin SC, Ye W, Fanshawe TR, Bullen C, Hartmann-Boyce J. Different doses, durations and modes of delivery of nicotine replacement therapy for smoking cessation. *Cochrane Database of Systematic Reviews*. 2019;4:CD013308.
38. Martínez-González MA, Buil-Cosiales P, Corella D, Bulló M, Fitó M, Vioque J, *et al.* PREDIMED-Plus Study Investigators. Cohort Profile: Design and methods of the PREDIMED-Plus randomized trial. *Int J Epidemiol*. 2019;48:387-388o.
39. Hernán MÁ, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clin Trials*. 2012;9:48-55.
40. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310:452-4.
41. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, *et al.* Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ*. 2020;192:E901-E906.
42. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, *et al.* RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.
43. Rothman KJ, Greenland S. Precision and validity in epidemiological studies. In: Rothman KJ, Greenland S. *Modern epidemiology*. 2nd ed. Lippincott-Raven Publishers. Philadelphia;1998:115-34.
44. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, *et al.* Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*. 2012;344:e1119.
45. Goldacre B, Drysdale H, Dale A, Milosevic I, Slade E, Hartley P, *et al.* COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*. 2019;20:118.
46. Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. *Cochrane Database Syst Rev*. 2012;12:MR000033.
47. Bes-Rastrollo M, Schulze MB, Ruiz-Canela M, Martínez-González MA. Financial conflicts of interest and reporting bias regarding the association between sugar-sweetened beverages and weight gain: a systematic review of systematic reviews. *PLoS Med*. 2013;10:e1001578.

48. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, *et al.* Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294:2203-9.
49. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294:218-28.
50. Heffner AC, Milbrandt EB, Venkataraman R. Trials stopped early for benefit? Not so fast! Crit Care. 2007;11:305.
51. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med. 2010;152:726-32.
52. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, *et al.* CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.
53. Frieden TR. Evidence for Health Decision Making - Beyond Randomized, Controlled Trials. N Engl J Med. 2017;377:465-75.

Anexo I. Lista de comprobación de la declaración SPIRIT 2013: elementos que se recomienda abordar en el protocolo de un ensayo clínico y en los documentos relacionados.

Sección/elemento	N.º	Descripción
INFORMACIÓN ADMINISTRATIVA		
Título	1	Título descriptivo que indique el diseño, la población, las intervenciones y, cuando corresponda, la sigla o abreviatura del ensayo.
Registro del ensayo	2a	Identificador del estudio y nombre del registro. Si no se lo ha registrado aun, nombre del registro donde se propone inscribirlo.
	2b	Todos los elementos del conjunto de datos del registro de ensayos de la Organización Mundial de la Salud, (véase el cuadro del apéndice, disponible en la versión original en inglés en www.annals.org).
Versión del protocolo	3	Fecha e identificación de la versión.
Financiación	4	Fuentes y tipo de apoyo financiero, material o de otra índole.
Funciones y responsabilidades	5a	Nombre, afiliaciones y funciones de quienes contribuyeron con el protocolo.
	5b	Nombre e información de contacto del patrocinador del estudio.
	5c	Funciones del patrocinador y de los financiadores del estudio, si los hubiera, en el diseño; la recolección, gestión, análisis e interpretación de los datos; la redacción de los informes; y la decisión de someterlo para publicación, especificando si tendrán la última palabra en cualquiera de estas actividades.
	5d	Composición, funciones y responsabilidades del centro coordinador, el comité directivo, el comité evaluador del criterio de valoración, el equipo de gestión de datos y de cualquier otra persona o grupo que supervise el ensayo, si corresponde (véase el elemento 21a sobre el comité de monitoreo de datos).
INTRODUCCIÓN		
Antecedentes y justificación	6a	Descripción de la pregunta de investigación y justificación para emprender el ensayo, incluido un resumen de los estudios relevantes (publicados y no publicados) que hayan analizado los beneficios y daños de cada intervención.
	6b	Explicación de la elección del comparador o los comparadores.
Objetivos	7	Objetivos o hipótesis específicos.
Diseño del ensayo	8	Descripción del diseño del ensayo, incluido el tipo de ensayo (por ejemplo, de grupos paralelos, de grupos cruzados, factorial, de un solo grupo), razón de asignación y marco de trabajo (por ejemplo, superioridad, equivalencia, no inferioridad, exploratorio).

Sección/elemento	N.º	Descripción
MÉTODOS		
Participantes, intervenciones y resultados		
Ámbito del estudio	9	Descripción de los centros del estudio (por ejemplo, consultorio de atención primaria, hospital académico) y lista de los países donde se recopilarán los datos. Referencia a dónde puede obtenerse la lista de los centros participantes.
Criterios de selección	10	Criterios de inclusión y de exclusión de los participantes. Si corresponde, criterios de selección de los centros del estudio y las personas que realizarán las intervenciones (por ejemplo, cirujanos o psicoterapeutas).
Intervenciones	11a	Intervenciones para cada grupo con detalles suficientes que permitan reproducirlas, incluido cómo y cuándo serán administradas.
	11b	Criterios para interrumpir o modificar las intervenciones asignadas a cada sujeto en el ensayo (por ejemplo, cambio en la dosis por daños al participante, a petición del participante o debido a una mejoría o a un empeoramiento de la enfermedad).
	11c	Estrategias para mejorar el cumplimiento de protocolos de intervención, así como cualquier método para vigilar el cumplimiento (por ejemplo, retorno de la medicación, pruebas de laboratorio).
	11d	Cuidados concomitantes e intervenciones relevantes permitidos y prohibidos durante el ensayo.
Resultados	12	Resultado principal y secundario, y otras valoraciones de la evolución o el desenlace, incluida la variable específica de medición (por ejemplo, presión arterial sistólica), la métrica de análisis (por ejemplo, cambio con respecto al valor inicial o línea de base, valor final, o tiempo hasta el evento), el método de agregación (por ejemplo, mediana, proporción) y el momento de cada variable. Se recomienda especialmente explicar la relevancia clínica de las variables de eficacia y seguridad (daños) seleccionadas.
Cronograma con respecto a los participantes	13	Cronograma para reclutar, realizar las intervenciones (incluidos períodos de preinclusión y de lavado), evaluar y visitar a los participantes. Se recomienda especialmente incluir un diagrama esquemático.
Tamaño de la muestra	14	Número estimado de participantes que se necesitan para alcanzar los objetivos del estudio y explicación sobre cómo se determinó dicho número, incluidas las premisas clínicas y estadísticas que respalden el cálculo del tamaño de la muestra.
Reclutamiento	15	Estrategias para lograr el reclutamiento adecuado a fin de alcanzar el tamaño de muestra previsto.

Sección/elemento	N.º	Descripción
Asignación de las intervenciones (para los ensayos clínicos controlados)		
Asignación generación de la secuencia	16a	Método para generar la secuencia de asignación (por ejemplo, números aleatorios generados por computadora) y lista de cualquier factor utilizado en la estratificación. Para reducir la predictibilidad de la secuencia aleatoria, los detalles de cualquier restricción planificada (por ejemplo, bloques) deberían facilitarse en un documento aparte que no esté disponible para quienes deben reclutar a los participantes o asignar las intervenciones.
Mecanismo para ocultar la asignación	16b	Mecanismo para aplicar la secuencia de asignación (por ejemplo, un teléfono central o sobres sellados, opacos y numerados de manera secuencial), con una descripción de toda medida para ocultar la secuencia hasta que se hayan asignado las intervenciones.
Ejecución	16c	Quién generará la secuencia de asignación, quién reclutará a los participantes y quién asignará las intervenciones a los participantes.
Enmascaramiento (cegamiento)	17a	Quién estará enmascarado después de la asignación de las intervenciones (por ejemplo, los participantes en el ensayo, los prestadores de atención de salud, los evaluadores de los resultados, los analistas de los datos) y cómo se hará el enmascaramiento.
	17b	Si hay enmascaramiento, circunstancias bajo las cuales se permite desenmascarar y procedimiento para revelar la intervención asignada a un participante durante el ensayo.
Recolección, gestión y análisis de datos		
Métodos de recolección de datos	18a	Planes para evaluar y recoger las variables iniciales, de evolución y otros datos del estudio, incluido cualquier proceso para mejorar la calidad de los datos (por ejemplo, mediciones por duplicado, capacitación de los evaluadores) y descripción de los instrumentos utilizados en el estudio (por ejemplo, cuestionarios, pruebas de laboratorio) junto con su fiabilidad y validez, si se conocen. Indicar dónde pueden encontrarse los formularios de recolección de datos, si no se encuentran en el protocolo.
	18b	Planes para promover la retención de los participantes y lograr un seguimiento completo, incluida una lista de los datos que se recopilarán de los participantes que abandonen el ensayo o se desvíen de él.
Gestión de datos	19	Planes para ingresar, codificar, proteger y guardar los datos, incluido cualquier proceso para mejorar su calidad (por ejemplo, ingreso por duplicado o revisión del rango de valores). Especificar dónde pueden encontrarse los detalles del procedimiento de gestión de datos que no figuren en el protocolo.
Métodos estadísticos	20a	Métodos estadísticos para analizar la variable principal y las secundarias. Especificar dónde pueden encontrarse los detalles del plan de análisis estadístico que no figuren en el protocolo.
	20b	Métodos para cualquier otro análisis adicional (por ejemplo, análisis de subgrupos o análisis ajustados).
	20c	Definición de la población de análisis en relación con la falta de cumplimiento del protocolo (por ejemplo, análisis conforme a la aleatorización) y de cualquier método estadístico para tratar los datos faltantes (por ejemplo, imputación múltiple).

Sección/elemento	N.º	Descripción
Monitoreo		
Monitoreo de datos	21a	Composición del comité de monitoreo de datos, resumen de su función y procedimiento de notificación, declaración sobre su independencia con respecto al patrocinador y sobre sus conflictos de intereses. Especificar dónde pueden encontrarse otros detalles sobre sus estatutos que no se hayan incluido en el protocolo. Alternativamente, explicar por qué no se necesita este comité.
	21b	Descripción de cualquier análisis intermedio y de las reglas de interrupción, incluido quién tendrá acceso a los resultados intermedios y quien tomará la decisión final de terminar el ensayo.
Daños	22	Planes para recoger, evaluar, informar y gestionar los eventos adversos, tanto los informados espontáneamente como aquellos cuya información se obtuvo al solicitarla, y otros efectos imprevistos de las intervenciones o de la realización del estudio.
Auditoría	23	Frecuencia y métodos para auditar la realización del estudio, si los hubiera, y si estos procesos son independientes de los investigadores y del patrocinador.
ÉTICA Y DISEMINACIÓN		
Aprobación ética de la investigación	24	Planes para obtener la aprobación del comité de ética o de la junta de revisión institucional.
Enmiendas del protocolo	25	Planes para comunicar las enmiendas importantes introducidas en el protocolo (por ejemplo, cambios en los criterios de selección, en las variables de resultados, en el análisis) a las partes pertinentes (por ejemplo, investigadores, comité de ética o junta de revisión institucional, participantes en el ensayo, registros de ensayos, revistas biomédicas, organismos reguladores).
Consentimiento o conformidad	26a	Quién obtendrá el consentimiento informado o el asentimiento de los participantes en el estudio o de sus representantes autorizados y cómo se hará (véase el elemento 32).
	26b	Disposiciones adicionales de consentimiento para la recolección y el uso de datos y muestras biológicas de los participantes de estudios auxiliares, cuando corresponda.
Confidencialidad	27	Cómo se recopilará, compartirá y guardará la información de carácter personal de los participantes potenciales y reclutados, con el fin de proteger la confidencialidad antes, durante y después del ensayo.
Declaración de intereses	28	Conflictos de interés de tipo económico o de otra índole de los investigadores principales tanto para el ensayo en su conjunto como de cada centro.
Acceso a los datos	29	Declaración de quién tendrá acceso al conjunto de datos finales del ensayo y revelación de los acuerdos contractuales que limiten tal acceso a los investigadores.

Sección/elemento	N.º	Descripción
Atención adicional y posterior al estudio	30	Disposiciones, si las hubiera, con respecto a la atención adicional y posterior al ensayo, y para la compensación en el caso de aquellos que sufran algún daño por haber participado en el estudio.
Política de diseminación	31a	Planes de los investigadores y del patrocinador para comunicar los resultados del ensayo a los participantes, los profesionales de la salud, el público y otros grupos pertinentes (por ejemplo, en una publicación, presentación de información en bases de datos de resultados u otros arreglos para difundir los datos), incluida cualquier restricción de publicación.
	31b	Pautas para elegir a los autores y si se pretende usar redactores científicos profesionales.
	31c	Planes, de haberlos, para hacer público el protocolo completo, el conjunto de datos de los participantes y el código estadístico.
APÉNDICES		
Materiales del consentimiento informado	32	Modelo del formulario del consentimiento informado y documentación relacionada que se entregue a los participantes o a sus representantes legales autorizados.
Muestras biológicas	33	Planes para recoger, estudiar y guardar muestras biológicas para análisis genéticos o moleculares, tanto presentes como futuros, cuando corresponda.

Fuente: Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 Statement: Defining standard protocol items for clinical trials [Spanish]. *Rev Panam Salud Publica*. 2015;38:506-14.

Anexo II. Lista de verificación TIDieR (plantilla para descripción de intervención y replicación)*: información a incluir al describir una intervención y la ubicación de la información.



Número de ítem	Ítem	Donde se encuentra **	
		Papel primario (número de página o apéndice)	Otros † (detalles)
	BREVE NOMBRE		
1.	Proporcione el nombre o una frase que describa la intervención.	_____	_____
	POR QUÉ		
2.	Describa cualquier razón, teoría u objetivo de los elementos esenciales para la intervención.	_____	_____
	QUÉ		
3.	Materiales: Describa cualquier material físico o informativo utilizado en la intervención, incluidos los proporcionados a los participantes o utilizados en la entrega de intervención o en la capacitación de proveedores de intervención. Proporcione información sobre dónde se puede acceder a los materiales (por ejemplo, apéndice en línea, URL).	_____	_____
4.	Procedimientos: Describa cada uno de los procedimientos, actividades y/o procesos utilizados en la intervención, incluidas las actividades de apoyo o habilitación.	_____	_____

Número de ítem	Ítem	Donde se encuentra **	
		Papel primario (número de página o apéndice)	Otros † (detalles)
	QUIÉN PROPORCIONÓ		
5.	Para cada categoría de proveedor de intervención (por ejemplo, psicólogo, asistente de enfermería), describa su experiencia, antecedentes y cualquier capacitación específica brindada.	_____	_____
	CÓMO		
6.	Describa los modos de provisión (por ejemplo, cara a cara o mediante algún otro mecanismo, como Internet o teléfono) de la intervención y si se aplicó individualmente o en grupo.	_____	_____
	DÓNDE		
7.	Describa los tipos de ubicaciones donde se produjo la intervención, incluida la infraestructura necesaria o las características relevantes.	_____	_____
	CUÁNDO y CUÁNTO		
8.	Describa la cantidad de veces que se realizó la intervención y durante qué período de tiempo, incluyendo la cantidad de sesiones, su horario y su duración, intensidad o dosis.	_____	_____
	PERSONALIZACIÓN		
9.	Si la intervención fue planeada para ser personalizada, titulada o adaptada, entonces describa qué, por qué, cuándo y cómo.	_____	_____
	MODIFICACIONES		
10.*	Si la intervención fue modificada durante el curso del estudio, describa los cambios (qué, por qué, cuándo y cómo).	_____	_____

Número de ítem	Ítem	Donde se encuentra **	
		Papel primario (número de página o apéndice)	Otros † (detalles)
	CÓMO DE BIEN		
11.	Planificada: si se evaluó la adherencia o la fidelidad a la intervención, describa cómo y por quién, y si se usaron estrategias para mantener o mejorar la fidelidad, descríbalas.	_____	_____
12.‡	Ejecutada: si se evaluó el cumplimiento o la fidelidad de la intervención, describa el grado en que la intervención se aplicó según lo planeado.	_____	_____

**** Autores: use N/A si un elemento no es aplicable para la intervención que se describe. Revisores: utilice “?” Si la información sobre el elemento no se informa / no se informa lo suficiente.**

† Si la información no se proporciona en el documento principal, proporcione detalles sobre dónde está disponible esta información. Esto puede incluir ubicaciones como un protocolo publicado u otros documentos publicados (proporcionar detalles de citas) o un sitio web (proporcionar la URL).

‡ Si completa la lista de verificación TIDieR para un protocolo, estos elementos no son relevantes para el protocolo y no pueden describirse hasta que se complete el estudio.

* Recomendamos encarecidamente utilizar esta lista de verificación junto con la guía TIDieR (ver BMJ 2014; 348: g 1687) que contiene una explicación y elaboración para cada ítem.

* El objetivo de TIDieR es informar detalles de los elementos de intervención (y, cuando sea relevante, elementos de comparación) de un estudio. Otros elementos y características metodológicas de los estudios están cubiertos por otras declaraciones de informes y listas de verificación y no se han duplicado como parte de la lista de verificación de TIDieR. Cuando se informa un ensayo aleatorio, la lista de verificación TIDieR debe usarse junto con la declaración CONSORT (ver www.consort-statement.org) como una extensión del Artículo 5 de la Declaración CONSORT 2010. Cuando se informa un protocolo de ensayo clínico, la lista de verificación TIDieR debe usarse junto con la declaración SPIRIT como una extensión del Artículo 11 de la Declaración SPIRIT 2013 (ver www.spirit-statement.org). Para diseños de estudio alternativos, TIDieR se puede usar junto con la lista de verificación apropiada para ese diseño de estudio (ver www.equator-network.org).

Anexo III. Consort 2010 lista de comprobación de la información que hay que incluir al comunicar un ensayo clínico aleatorizado*



Sección/tema	Item n.º	Ítem de la lista de comprobación	Informado en página n.º
Título y resumen			
	1a	Identificado como un ensayo aleatorizado en el título.	
	1b	Resumen estructurado del diseño, métodos, resultados y conclusiones del ensayo (para una orientación específica, véase "CONSORT for abstracts").	
Introducción			
Antecedentes y objetivos	2a	Antecedentes científicos y justificación.	
	2b	Objetivos específicos o hipótesis.	
Métodos			
Diseño del ensayo	3a	Descripción del diseño del ensayo (p. ej., paralelo, factorial), incluida la razón de asignación.	
	3b	Cambios importantes en los métodos después de iniciar el ensayo (p. ej., criterios de selección) y su justificación.	
Participantes	4a	Criterios de selección de los participantes.	
	4b	Procedencia (centros e instituciones) en que se registraron los datos.	
Intervenciones	5	Las intervenciones para cada grupo con detalles suficientes para permitir la replicación, incluidos cómo y cuándo se administraron realmente.	
Resultados	6a	Especificación a priori de las variables respuesta (o desenlace) principal(es) y secundarias, incluidos cómo y cuándo se evaluaron.	
	6b	Cualquier cambio en las variables respuesta tras el inicio del ensayo, junto con los motivos de la(s) modificación(es).	

Sección/tema	Item n.º	Ítem de la lista de comprobación	Informado en página n.º
Tamaño muestral	7a	Cómo se determinó el tamaño muestral.	
	7b	Si corresponde, explicar cualquier análisis intermedio y las reglas de interrupción.	
Aleatorización			
Generación de la secuencia	8a	Método utilizado para generar la secuencia de asignación aleatoria.	
	8b	Tipo de aleatorización; detalles de cualquier restricción (como bloques y tamaño de los bloques).	
Generación de la secuencia	9	Mecanismo utilizado para implementar la secuencia de asignación aleatoria (como contenedores numerados de modo secuencial), describiendo los pasos realizados para ocultar la secuencia hasta que se asignaron las intervenciones.	
Implementación	10	Quién generó la secuencia de asignación aleatoria, quién seleccionó a los participantes y quién asignó los participantes a las intervenciones.	
Enmascaramiento	11a	Si se realizó, a quién se mantuvo cegado después de asignar las intervenciones (p. ej., participantes, cuidadores, evaluadores del resultado) y de qué modo.	
	11b	Si es relevante, descripción de la similitud de las intervenciones.	
Métodos estadísticos	12a	Métodos estadísticos utilizados para comparar los grupos en cuanto a la variable respuesta principal y las secundarias.	
	12b	Métodos de análisis adicionales, como análisis de subgrupos y análisis ajustados.	
Resultados			
Flujo de participantes (se recomienda encarecidamente un diagrama de flujo)	13a	Para cada grupo, el número de participantes que se asignaron aleatoriamente, que recibieron el tratamiento propuesto y que se incluyeron en el análisis principal.	
	13b	Para cada grupo, pérdidas y exclusiones después de la aleatorización, junto con los motivos.	
Reclutamiento	14a	Fechas que definen los períodos de reclutamiento y de seguimiento.	
	14b	Causa de la finalización o de la interrupción del ensayo.	

Sección/tema	Ítem n.º	Ítem de la lista de comprobación	Informado en página n.º
Datos basales	15	Una tabla que muestre las características basales demográficas y clínicas para cada grupo.	
Números analizados	16	Para cada grupo, número de participantes (denominador) incluidos en cada análisis y si el análisis se basó en los grupos inicialmente asignados.	
Resultados y estimación	17a	Para cada respuesta o resultado final principal y secundario, los resultados para cada grupo, el tamaño del efecto estimado y su precisión (como intervalo de confianza del 95 %).	
	17b	Para las respuestas dicotómicas, se recomienda la presentación de los tamaños del efecto tanto absoluto como relativo.	
Análisis secundarios	18	Resultados de cualquier otro análisis realizado, incluido el análisis de subgrupos y los análisis ajustados, diferenciando entre los especificados a priori y los exploratorios.	
Daños (Perjuicios)	19	Todos los daños (perjuicios) o efectos no intencionados en cada grupo (para una orientación específica, véase "CONSORT for harms").	
Discusión			
Limitaciones	20	Todos los daños (perjuicios) o efectos no intencionados en cada grupo (para una orientación específica, véase "CONSORT for harms").	
Generalización	21	Posibilidad de generalización (validez externa, aplicabilidad) de los hallazgos del ensayo.	
Interpretación	22	Interpretación consistente con los resultados, con balance de beneficios y daños, y considerando otras evidencias relevantes.	
Otra información			
Registro	23	Número de registro y nombre del registro de ensayos	
Protocolo	24	Dónde puede accederse al protocolo completo del ensayo, si está disponible.	
Financiación	25	Fuentes de financiación y otras ayudas (como suministro de medicamentos), papel de los financiadores.	

* Recomendamos de modo encarecido leer esta lista de comprobación junto con "the CONSORT 2010 Explanation and Elaboration" para aclarar dudas importantes sobre todos los ítems. Si procede, también recomendamos leer las extensiones de CONSORT para ensayos aleatorizados por conglomerados, ensayos de no inferioridad y equivalencia, tratamientos no farmacológicos, intervenciones de medicamentos herbales y ensayos pragmáticos. Se están preparando otras extensiones: para éstas y para referencias actualizadas relevantes, relacionadas con esta lista de verificación, véase www.consort-statement.org

Capítulo 10. Estudios de cohortes

MARINA POLLÁN SANTAMARÍA y BEATRIZ PÉREZ GÓMEZ

1. Algunos ejemplos de estudios de cohortes
2. Clasificación de los estudios de cohortes
3. Ventajas e inconvenientes de los estudios de cohortes frente a los estudios de casos y controles
4. Aspectos básicos del diseño de los estudios de cohortes
5. Selección de la cohorte
6. Fechas de entrada y salida: contribución de cada miembro de la cohorte al estudio
7. Seguimiento
8. Eventos finales de interés
9. Valoración de la exposición
10. La estimación de la exposición
11. Otros factores de riesgo: confusión e interacción
12. Confidencialidad
13. Estrategias de análisis
14. Potencia estadística
15. Problemas de interpretación de los estudios de cohortes
16. Casos y controles basados en estudios de cohortes
17. Estudios de cohortes y los ensayos clínicos
18. Conclusiones
19. Referencias

Los estudios de cohortes se caracterizan por la identificación y seguimiento de uno o más grupos de individuos con un determinado nivel de exposición, para detectar y cuantificar la aparición del evento o enfermedad de interés a lo largo del tiempo.

La palabra “cohorte” deriva del vocablo latino *cohors-cohortis*, que designaba a una unidad táctica de infantería dentro de las legiones romanas. En epidemiología, el término “cohorte” fue incorporado en 1935 y sirve en la actualidad para referirse a un conjunto de personas que comparten una determinada exposición y/o “evolucionan” juntas a lo largo del tiempo.

La esencia de los estudios de cohortes es el seguimiento. En el diseño de cohortes está implícita la recogida de información, a nivel individual, sobre la exposición de uno o varios grupos de personas (cohortes), y el establecimiento de un mecanismo para seguir a los miembros de estas cohortes con objeto de detectar la aparición del evento de interés y también la posible salida o abandono del estudio. El seguimiento sirve para cuantificar no sólo la frecuencia de enfermedad en las distintas cohortes (incidencia acumulada), sino el ritmo o dinámica de la enfermedad en cada una de ellas (tasa de incidencia). Además, durante el seguimiento es posible considerar e incorporar en el análisis los cambios en el nivel de exposición así como las variaciones del resto de factores de riesgo relevantes.

Por tanto, los estudios de cohortes valoran el estado de exposición antes de que se produzca la enfermedad y permiten registrar los cambios ocurridos en el tiempo a nivel individual. La mayor parte de las enfermedades crónicas son el resultado de un proceso que se extiende a lo largo de décadas, en el que intervienen diferentes exposiciones o factores de riesgo. La naturaleza dinámica de la exposición a estos factores y su interrelación a lo largo del tiempo sólo puede ser estudiada adecuadamente en un diseño de cohortes. Esta dinámica temporal está ausente de los estudios transversales y es difícilmente investigable en los estudios de casos y controles.

1. Algunos ejemplos de estudios de cohortes

A continuación se presentan 6 estudios de cohortes que servirán de ejemplo para ilustrar los aspectos básicos del diseño desarrollados más adelante.

Ejemplo 1: Cohorte de médicos británicos:

El estudio prospectivo de los efectos del tabaco sobre la salud comenzó en 1951. En 1964 los autores publicaron un artículo con el diseño de su estudio y los resultados correspondientes a los primeros diez años de seguimiento. El estudio invitaba a participar a los médicos de Reino Unido, rellenando un cuestionario que fue enviado el 31 de octubre de 1951 a 59 600 médicos de ambos sexos que figuraban en el Registro de Médicos. Los investigadores optaron por un cuestionario corto que favoreciese la participación. Se pidió a los médicos que se clasificasen como: fumadores actuales, exfumadores o no fumadores (definidos como aquellos que no han llegado a fumar un cigarrillo al día durante 1 año). A los fumadores o exfumadores se les pidió que informasen sobre la edad de inicio y la cantidad de cigarrillos/día que fumaban en ese momento o antes de abandonar el hábito.

Un total de 40 637 médicos (34 445 hombres y 6192 mujeres) contestaron el cuestionario inicial. Durante el seguimiento se les enviaron otros tres cuestionarios a los hombres y otros dos a las mujeres para estimar los cambios en el patrón de consumo.

El seguimiento se realizó mediante el Registro Nacional de Defunción. En 1964 se publicaron los resultados de los diez primeros años de seguimiento y posteriormente los correspondientes a veinte y veintidós años de seguimiento en hombres y en mujeres respectivamente. Se contabilizaron 467 muertes por cáncer de pulmón como causa básica de defunción, y otras 20 citaban dicha enfermedad como causa contribuyente. Para estas 487 muertes se buscó confirmación del diagnóstico a partir del médico que rellenó el certificado de defunción. Como evento de interés se estudió también la mortalidad por otros tumores, enfermedades respiratorias, enfermedades cardiovasculares y otras.

Ejemplo 2: Estudio Prospectivo Europeo de Dieta y Cáncer (EPIC):

El EPIC (European Prospective Investigation into Cancer and Nutrition) es un estudio prospectivo multinacional coordinado desde la Agencia Internacional de Investigación del Cáncer (IARC). Este estudio ha sido diseñado para investigar la relación entre diferentes factores nutricionales y la incidencia de cáncer y otras enfermedades crónicas. Se han reclutado 520 000 participantes en diez países europeos: Alemania, Dinamarca, España, Francia, Grecia, Holanda, Italia, Noruega, Reino Unido y Suecia. El carácter internacional del estudio aumenta la heterogeneidad dietética de los participantes, lo que permite comparar patrones de alimentación muy diversos. El reclutamiento se inició en 1992. Aunque, por lo general, la fuente de reclutamiento de los participantes ha sido la población general, la forma de seleccionar y seguir a los participantes varía de unos países a otros. La información sobre dieta se obtuvo mediante un cuestionario dietético que fue adaptado y validado en todos los países participantes. Se obtuvieron también muestras de sangre, que fueron almacenadas para análisis posteriores. El estudio se plantea un seguimiento mínimo de 10 años. Los casos de cáncer que se van produciendo en la cohorte son detectados mediante los registros de cáncer poblacionales de las áreas del estudio. España contribuye en este estudio con 41 440 participantes, reclutados por los Registros de Cáncer de Granada, Murcia, Oviedo, Pamplona y San Sebastián. El centro coordinador español está en Barcelona.

Ejemplo 3: El estudio de las enfermeras americanas NHS (Nurses' Health Study)

El primer estudio se inició en 1976, y su objetivo principal era investigar los efectos a largo plazo del consumo de anticonceptivos. Se eligieron las enfermeras debido a su alto nivel educativo y a su capacidad para contestar las preguntas relevantes del estudio. El diseño inicial se centró en la población de enfermeras casadas, con edades entre 30 y 55 años en 1976 que vivían en 11 estados de USA, en los que fue posible obtener los listados de las enfermeras y su dirección. Se consideró

incluidas en la cohorte a todas las enfermeras que contestaron el cuestionario inicial, (122 000 enfermeras de un total de 170 000 a las que se había enviado dicho cuestionario).

Cada dos años las enfermeras de la cohorte han de rellenar un cuestionario de seguimiento incluyendo preguntas sobre su salud, consumo de tabaco, uso de anticonceptivos, hormonas y estatus menopáusico.

A partir de 1980 se incluyeron cuestionarios dietéticos con una periodicidad cuatrienal. Además, para tener en cuenta aspectos dietéticos difícilmente investigables a través de cuestionario, se pidió a las enfermeras que enviaran cortes de las uñas de los pies. En 1989 se recogieron muestras de sangre de una submuestra de la cohorte (32 826 mujeres). Finalmente en 1992 se añadió un cuestionario de calidad de vida, que se cumplimenta también cada cuatro años.

Al NHS-I se ha sumado el NHS-II en 1989, cuyo objetivo es estudiar el efecto de los anticonceptivos orales, la dieta y los estilos de vida en una población más joven, con edades comprendidas entre los 25 y los 42 años al inicio del estudio. Esta generación incluye mujeres que empezaron a tomar anticonceptivos durante la adolescencia y por ello han tenido una mayor exposición a lo largo de la vida. En esta cohorte se han incluido casi 117 000 enfermeras.

Ejemplo 4: La cohorte MACS (Multicentric Aids Cohort Study)

Esta cohorte multicéntrica americana se constituyó para investigar la historia natural de la infección por VIH. EN 1984 se reclutaron 4954 hombres homosexuales en Baltimore, Chicago, Los Ángeles y Pittsburg, a los que se sumaron otros 668 varones entre 1987-1991, fundamentalmente para representar a las minorías no suficientemente consideradas en la muestra inicial (minorías étnicas y usuarios de drogas parenterales).

En el momento del reclutamiento inicial, se sabía que el SIDA era una enfermedad infecciosa, pero no se había descubierto el agente causal ni se disponía de las técnicas serológicas para su determinación. Cuando dichas técnicas estuvieron disponibles, el análisis de las muestras de sangre almacenadas de forma sistemática cada 6 meses permitió determinar el estatus serológico de los participantes. Se comprobó que el 39% de los participantes eran seropositivos en el momento del reclutamiento (seroprevalentes), mientras que entre los 3427 pacientes seronegativos, entre 1984 y 1997 se observaron 551 seroconversiones. Esta cohorte ha permitido investigar diferentes aspectos de la historia natural de la enfermedad, incluyendo la prevalencia y la incidencia del VIH, la inmunosupresión, el periodo de incubación, los marcadores de progresión de la enfermedad, la supervivencia desde el diagnóstico de SIDA, los factores de riesgo de la progresión rápida en los seropositivos, la efectividad de las nuevas terapias, etc.

Ejemplo 5: La cohorte de población activa sueca

En los países nórdicos, la existencia de registros poblacionales de gran calidad ha permitido la reconstrucción de cohortes históricas mediante el procedimiento de enlace de registros (record-linkage). Un ejemplo de este tipo de estudios es la cohorte de población activa sueca, identificada a través del censo de población de ese país de 1970, que ha sido seguida un total de 19 años (1971-1989) mediante enlace de registros, utilizando los registros nacionales suecos de mortalidad y de cáncer. La cohorte incluyó únicamente a los trabajadores de 1970 que estuviesen ya presentes en el país, aunque no trabajasen, en 1960. El enlace de registros se realizó utilizando el número de identificación personal, un código de 10 dígitos que identifica de forma inequívoca a cada ciudadano sueco.

La cohorte de trabajadores con edades comprendidas entre los 25 y los 64 años al inicio del seguimiento, incluyó 1 779 646 hombres y 1 066 346 mujeres. En el estudio inicial se presentaba la asociación entre la ocupación desempeñada en 1970 y la incidencia de 30 tipos de cáncer. De forma más específica, se ha investigado la incidencia de cáncer de mama y posteriormente de otras localizaciones.

Frente a otras cohortes históricas similares, la cohorte sueca presenta las siguientes peculiaridades: 1) estudio de ambos sexos y no sólo de los trabajadores varones, 2) cálculo exacto de la contribución de personas-año de cada participante del estudio, y 3) consideración de la ocupación referida en el censo de 1960 para definir subcohortes más específicas para cada ocupación, incluyendo sólo a los trabajadores clasificados bajo ese código ocupacional en ambos censos.

Ejemplo 6: Cohortes hospitalarias: Estudio de factores pronósticos de cáncer de pulmón no microcítico

Los estudios de factores pronósticos frecuentemente utilizan cohortes hospitalarias de pacientes. El ejemplo que planteamos es un estudio multicéntrico sobre factores pronósticos de pacientes con cáncer de pulmón no microcítico (CPNM) reseccable, llevado a cabo de forma prospectiva en 13 hospitales españoles. En el estudio se incluyeron a todos los pacientes con cáncer de pulmón no microcítico, intervenidos en los centros participantes, que cumpliesen los criterios de inclusión: casos con tumor reseccable con confirmación histológica, no fallecido como consecuencia directa del procedimiento quirúrgico y sin quimioterapia o radioterapia pre-quirúrgica. En total, el estudio incluyó 465 pacientes.

El objetivo del estudio era evaluar el posible papel como factor pronóstico de determinados biomarcadores: niveles séricos preoperatorios de algunos antígenos relacionados con el cáncer como el antígeno carcinoembrionario (CEA) y CA 125, y la sobre-expresión en el tejido tumoral de algunas proteínas como p53 y c-erbB-2. Todas las determinaciones se hicieron de forma centralizada para minimizar la variabilidad debida al laboratorio. Además, en aquellos pacientes que aceptaron participar, se obtuvo información detallada sobre el consumo de tabaco y la historia ocupacional completa. Se investigó también el posible efecto pronóstico del consumo de tabaco, utilizando la información procedente de los últimos 5 años previos al diagnóstico.

El momento de la operación quirúrgica supone el inicio del seguimiento. Los pacientes fueron citados con un intervalo mínimo de 6 meses entre las visitas médicas. Los eventos de interés fueron la muerte por cáncer de pulmón y la aparición de recidiva. También se utilizó el teléfono para localizar a los pacientes que no acudían al hospital y confirmar su estado de salud. Algo original en este estudio ha sido la investigación de la relación entre los factores pronósticos y el tipo de recidiva, diferenciando entre recidiva locorregional y aparición de metástasis a distancia.

2. Clasificación de los estudios de cohortes

Considerando la referencia temporal del investigador respecto del seguimiento, los estudios de cohortes se clasifican en:

- **Estudios prospectivos:** El diseño se realiza a priori, de forma que todo el seguimiento de los miembros de la(s) cohorte(s) se sitúa cronológicamente dentro del periodo del estudio. El estudio de los médicos británicos, el EPIC, el estudio de las enfermeras (NHS) y el estudio de factores pronósticos de CPNM son estudios prospectivos.
- **Estudios retrospectivos o cohortes históricas:** El estudio se inicia a posteriori. Se trata de reconstruir el seguimiento de los miembros de la cohorte a través de la información recogida en registros históricos (ej: censos, padrones, registros de cáncer, historias clínicas, etc.). Lógicamente en estos estudios no será posible obtener información que no esté recogida en los registros utilizados. Por ello, el factor limitante es la calidad de las fuentes de información utilizadas. El estudio de la cohorte ocupacional sueca es un ejemplo de cohorte histórica.
- **Estudios ambispectivos:** Es una mezcla de los dos anteriores. En el momento de iniciar el estudio, parte del seguimiento de todos o de algunos de los miembros de la cohorte ya ha transcurrido, por lo que se hace necesario incorporar esta información histórica al resto de los datos recogidos de forma prospectiva. Los estudios de cohortes ocupacionales frecuentemente son ambispectivos. Cuando se decide iniciar el estudio, algunos de los miembros de la cohorte llevan bastante tiempo trabajando en la industria de interés, mientras que los nuevos contratados se incorporan al estudio de forma totalmente prospectiva. El estudio MACS es un ejemplo mixto. Con respecto a la seroconversión, el estudio es prospectivo en los seronegativos y retrospectivo en los positivos.

Por otra parte, en función del grupo de comparación o referencia utilizado se pueden considerar tres tipos de estudios:

- **Estudios con dos cohortes: expuesta y no expuesta:** Una vez definida la cohorte expuesta, los investigadores seleccionan e incorporan en el estudio una cohorte no expuesta.
- **Estudios de comparación con la población general:** El estudio identifica y sigue a una determinada cohorte de exposición. La frecuencia de aparición de la enfermedad en dicha cohorte es comparada con la observada en la población general. Para ello es necesario disponer de registros poblacionales que suministren dicha información.
- **Comparaciones internas:** La cohorte a estudio supone una mezcla de personas expuestas y no expuestas. Las comparaciones se realizan dentro de la propia cohorte.

3. Ventajas e inconvenientes de los estudios de cohortes frente a los estudios de casos y controles

Frente a los estudios de casos y controles, los estudios de cohortes presentan las siguientes **ventajas**:

- Permiten obtener información más detallada de los efectos de una determinada exposición. Posibilitan la inclusión en el estudio de uno o más eventos de interés relacionados con dicha exposición.
- Presentan menor propensión a los sesgos de recuerdo y de selección. Por un lado, la relación exposición-enfermedad es estudiada respetando la secuencia temporal. La valoración de la exposición es anterior al evento de interés. Por otro lado, la población origen del estudio está mejor definida y es fácilmente identificable. El sesgo de selección es el punto crítico de los estudios de casos y controles, cuya validez radica en la adecuada inclusión de controles realmente representativos de la población origen de los casos.
- Facilitan el estudio de exposiciones infrecuentes, difíciles de incluir en estudios de casos y controles.
- Favorecen un uso adecuado de las muestras biológicas, que son recogidas antes de que aparezca la enfermedad.
- Posibilitan la medición de la exposición de forma repetida en el tiempo. Con ello permiten valorar los cambios que se producen en la exposición y también calibrar los posibles errores de medición.
- Permiten estudiar la incidencia de la enfermedad en la población a estudio y también el riesgo (incidencia acumulada).

Entre los **inconvenientes** de los estudios de cohortes, respecto a los casos y controles, figuran los siguientes:

- En general requieren mucho más tiempo. El seguimiento puede durar muchos años, lo que encarece el estudio y dificulta la retención de los participantes. Los estudios de cohortes históricas no sufren dicho inconveniente, pero su calidad está determinada por las fuentes de información utilizadas.
- Cuando el evento de interés es infrecuente (enfermedades de baja incidencia) es necesario incluir en el estudio un número elevado de participantes.
- Los estudios de cohortes basadas en grupos específicos de exposición, como pueden ser determinados sectores ocupacionales, no permiten estimar directamente la fracción de enfermedad atribuible a la exposición. Para ello, es necesario conocer cual es la frecuencia de exposición en la población general.

Vistos los pros y contras, podemos concluir que los estudios de cohortes y los caso-control son dos herramientas complementarias. Los estudios de cohortes proporcionan una población bien definida en la que identificar los casos de forma no sesgada y con información disponible de la secuencia temporal del seguimiento. Los caso-control concentran el esfuerzo en los “individuos informativos”. Por eso muchos estudios de cohortes incorporan la estrategia de utilizar diseños de caso-control dentro de la cohorte.

4. Aspectos básicos del diseño de los estudios de cohortes

A la hora de diseñar un estudio de cohortes es importante definir de forma clara los siguientes aspectos:

1. Selección de la cohorte: Criterios de inclusión.
2. Fecha de entrada y de salida de los participantes en el estudio.
3. Seguimiento.
4. Eventos finales de interés.
5. La exposición.
6. Factores de confusión.
7. Confidencialidad.
8. Estrategias de análisis.
9. Potencia estadística.
10. Optimización de recursos: estudios de caso-control sobre una cohorte.

5. Selección de la cohorte

La selección de la(s) cohorte(s) viene motivada por la naturaleza de la exposición, las características de la enfermedad estudiada y la posibilidad de seguimiento del grupo elegido.

Respecto a la exposición, es importante que la cohorte elegida incluya personas con un grado de exposición suficiente como para desarrollar el/los eventos de interés. En este sentido, el NHS-II incluye un colectivo nuevo de enfermeras en las que la posible exposición a hormonas contraceptivas ha sido mucho mayor que la de la cohorte inicial. Por otra parte, si el estudio pretende establecer comparaciones internas dentro de la cohorte, ésta debe ser suficientemente heterogénea en cuanto a niveles de exposición. Así, el carácter multinacional del estudio EPIC permitirá investigar la influencia de patrones dietéticos que normalmente presentan poca variabilidad en una región o incluso en un país.

Respecto al evento o enfermedad de interés, cohortes como el MACS fueron diseñadas para investigar los factores de riesgo de enfermedades de origen desconocido. En este sentido, interesa incluir en la(s) cohorte(s) colectivos con una elevada incidencia. Por este motivo, el reclutamiento inicial en el estudio MACS incluyó 5000 varones homosexuales, colectivo que sufrió el primer impacto de la epidemia de SIDA en Estados Unidos.

Teniendo en cuenta el seguimiento, es importante que el colectivo sea fácil de seguir a lo largo del estudio y que el porcentaje de abandonos del estudio (pérdidas esperadas) sea bajo. Ello ha motivado la creación de cohortes de médicos y de enfermeras para investigar exposiciones no relacionadas directamente con su trabajo, como son el consumo de tabaco y de anticonceptivos orales. Estos colectivos tienen un mayor grado de motivación y están, además, altamente cualificados a la hora de proporcionar información sobre la exposición e incluso sobre los eventos de interés.

Finalmente, la principal ventaja de cohortes históricas como la cohorte de población activa sueca es su carácter poblacional, lo que elimina los sesgos de selección y permite estimar la incidencia de cáncer y el riesgo relativo en ocupaciones de baja frecuencia.

La base poblacional de un estudio de cohortes se define como la experiencia de personas-tiempo incluida en ese estudio concreto. Los resultados del mismo se refieren precisamente a esa base poblacional. Para que dicha base poblacional sea fácilmente identificable, las reglas de inclusión y exclusión deben ser claras.

6. Fechas de entrada y salida: contribución de cada miembro de la cohorte al estudio

La fecha de entrada supone el inicio de la contribución de cada miembro de la cohorte al estudio, en términos de personas año. La fecha de entrada puede, y de hecho suele, no coincidir con el inicio de la exposición, ni con la fecha en la que dicha persona ha completado lo que podríamos considerar como “dosis mínima” de exposición relevante.

La fecha de salida está condicionada por tres factores:

1. El desarrollo del evento o enfermedad de interés, en caso de que dicho evento sea único;
2. el cierre del estudio, y
3. el abandono del estudio por la causa que sea.

A la hora de establecer el nivel de exposición, es importante considerar la información anterior a la inclusión del sujeto en el estudio. Sin embargo, ese periodo no debe contribuir en el cómputo de personas año. Dicha inclusión supondría un sesgo de selección, ya que el diseño del estudio ha dejado fuera a aquellos miembros de la población objeto del estudio que ya no están presentes (no son reclutables) al inicio del mismo, ya sea por haber fallecido, haber desarrollado la enfermedad previamente o haber abandonado dicha población.

En la tabla 1 se presenta la información sobre la entrada y salida y el tiempo biológicamente relevante en cada uno de los ejemplos citados al inicio de este capítulo.

Tabla 1. Periodo biológicamente relevante y periodo de estudio en los estudios de cohortes citados

Estudio	Periodo biológicamente relevante (PBR)	Entrada en el estudio	¿Incluye el estudio todo el PBR?
Médicos británicos	Del inicio del consumo de tabaco a la muerte por cáncer de pulmón u otras causas	Primer cuestionario	No
EPIC	Del inicio de consumo de determinado nutriente a la aparición de un det. tumor	Primer cuestionario y toma de muestras biológicas	No
NHS	Del inicio del consumo de anticonceptivos a la aparición de uno de los eventos de interés	Primer cuestionario	No
MACS	Seroconversión - desarrollo de SIDA y posteriormente muerte	Primer cuestionario y muestra de sangre	Sólo en sujetos seronegativos

Tabla 1. Periodo biológicamente relevante y periodo de estudio en los estudios de cohortes citados
(continuación)

Estudio	Periodo biológicamente relevante (PBR)	Entrada en el estudio	¿Incluye el estudio todo el PBR?
Cohorte sueca	Del inicio de una determinada ocupación a la aparición de un determinado tumor	1-Enero-1971	No
Factores pronósticos CPNM	Cirugía - recidiva tumoral - muerte	Fecha cirugía	Si

7. Seguimiento

La calidad del seguimiento es el principal factor para definir la calidad de un estudio de cohortes. En general, los objetivos del seguimiento son tres:

- ♦ Identificar la aparición del evento.
- ♦ Identificar las pérdidas del estudio y el momento en que se producen.
- ♦ Identificar cambios relevantes en la exposición y, si es necesario, en otros factores de riesgo relevantes.

El tipo de seguimiento es diferente según los estudios. En algunos, el seguimiento es pasivo, basado exclusivamente en la información disponible en los registros poblacionales, como el caso de los médicos británicos o la cohorte sueca. En otros, los sujetos participantes deben acudir de forma periódica al hospital o centro correspondiente, como es el caso del MACS y del estudio de factores pronósticos de CPNM. En otros casos, el seguimiento se realiza por correo, a través de encuestas (ejemplo: NHS).

La periodicidad del seguimiento viene determinada por la dinámica de la enfermedad y la necesidad de mantener a los participantes dentro del estudio. Aunque las técnicas de análisis estadístico permiten considerar la información parcial proporcionada por las personas que abandonan el estudio, un alto porcentaje de abandonos o pérdidas puede comprometer seriamente la generabilidad de los resultados. Por ello, muchos estudios de cohortes han desarrollado estrategias para mantener la tasa de participación de los sujetos incluidos (ver tabla 2).

Tabla 2. Estrategias para maximizar la adherencia

En el reclutamiento: • Investigar la intención de permanecer en la cohorte • Hacer que el participante complete con éxito las tareas que se requerirán de él en el seguimiento • Informar ampliamente de los requerimientos de su participación en el estudio • Recoger información para el seguimiento: dirección, teléfono, número SS, lugar y fecha de nacimiento, etc • Recoger información similar de contactos
Identificación con el estudio: • Identificación con el estudio: Crear un logo y un lema • Mandar cartas, felicitaciones, información sobre la marcha del estudio, etc.
Frecuencia de contacto: • Contactos regulares cada 6-24 meses • Recoger como mínimo información sobre el evento • Diseñar un sistema de rastreo
Responsables del seguimiento: • Entrenados y entusiastas • Que faciliten la comunicación • Que sean flexibles en la citación de los participantes
Uso de incentivos • Distribuir entre los participantes objetos con el logotipo del estudio • Cubrir los gastos de los envíos por correo • Informar periódicamente a los participantes de los resultados del estudio

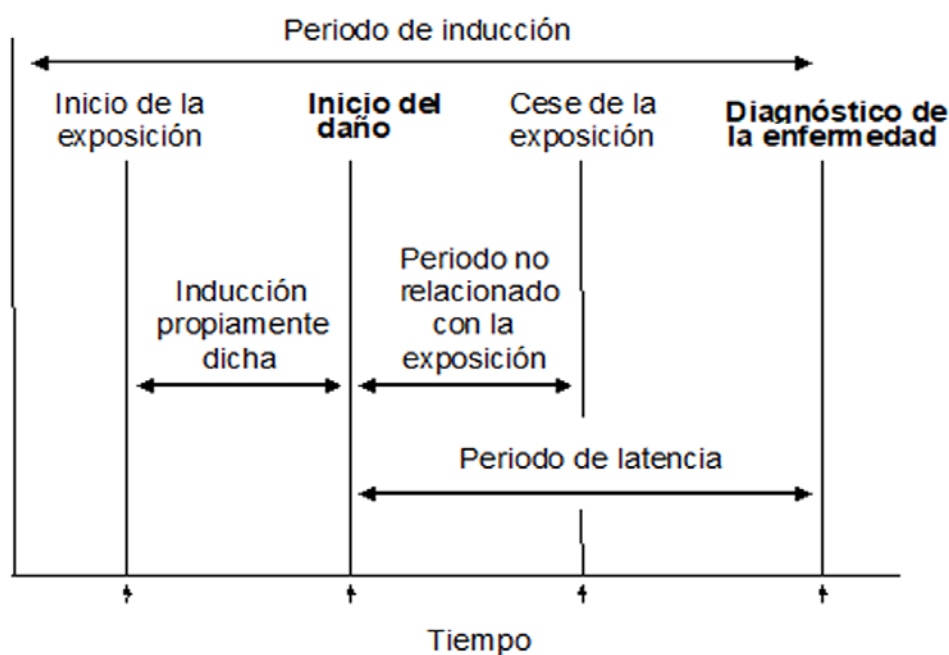
Es importante que el seguimiento cuente con un dispositivo capaz de detectar los abandonos del estudio y el momento en que se producen. En situaciones de seguimiento pasivo, cuando el diagnóstico de la enfermedad se realiza a través de registros, el investigador puede estar considerando dentro del estudio a sujetos que de hecho están ya fuera de él, lo que se traduce en una subestimación de la incidencia real. Por ejemplo, la cohorte de población activa sueca fue seguida mediante los registros suecos de cáncer y de defunción. Todas las personas que refirieron una ocupación concreta en 1970 se consideraron dentro del estudio mientras no figurase su fecha de defunción en el registro de mortalidad. Sin embargo, algunos trabajadores han podido salir del país y por ello abandonar el estudio. En este caso concreto, la imprecisión cometida al considerar a estas personas dentro de la cohorte es pequeña, dado que la tasa de emigración en Suecia en los años 70 y 80 fue menor del 1 por mil.

8. Eventos finales de interés

Durante el seguimiento es necesario disponer de un sistema que identifique en qué momento se produce el evento o enfermedad a estudio. En estudios de mortalidad, como la cohorte de médicos británicos, el momento en que se produce el evento es fácil de constatar. Lo que resulta más problemático en estos estudios es considerar la causa de defunción. La calidad de los certificados de defunción es heterogénea según el tipo de enfermedad considerada. En el caso de la mortalidad por cáncer, diferentes estudios han mostrado que la información contenida en los certificados de defunción es razonablemente válida.

En los estudios de incidencia basados en registros poblacionales, la fecha del evento es la fecha de diagnóstico que figura en el registro. En el estudio MACS, como fecha de seroconversión se consideró la media entre la última visita con serología negativa y la primera con serología positiva. En el estudio de factores pronósticos de CPNM, se consideró como fecha de recidiva la primera visita o exploración positiva.

La mayor parte de enfermedades crónicas plantean la necesidad de periodos extensos de seguimiento, debido a la magnitud de los periodos de inducción y latencia. Aunque algunos autores usan ambos términos como sinónimos, el periodo entre la exposición y la aparición de los síntomas de la enfermedad puede dividirse en dos partes (ver figura 1). El periodo de inducción es el intervalo de tiempo necesario para que el agente causal (sólo o con la concurrencia de otros factores) sea suficiente para iniciar el daño. A partir de este punto comenzaría el periodo de latencia, durante el cual la enfermedad progresa hasta el momento en el que es diagnosticable. Este periodo se reduce, por ejemplo, con programas de screening que detectan la enfermedad presintomáticamente. Si, una vez que se inicia el proceso patológico, la enfermedad puede revertir sin haber sido diagnosticada, ambos periodos se solapan; si éste no es el caso, el periodo de inducción y el de latencia son consecutivos. Tras el inicio de la enfermedad, la exposición puede seguir existiendo: en algunos casos la presencia del agente ya no tiene efecto sobre la evolución de la misma, mientras que en otros puede potenciar su progreso (agente promotor).

Figura 1. Relación temporal entre exposición y enfermedad.

Actualmente se investiga la posibilidad de utilizar biomarcadores como eventos intermedios o marcadores de efecto precoz, que permitirían disminuir la duración del estudio e instaurar medidas preventivas antes de que la enfermedad avance (ver figura 4). Por ejemplo, en el estudio de la historia natural de infección por VIH, la caída de linfocitos CD4⁺ es un determinante de progresión de enfermedad. Para las enfermedades cardiovasculares se han utilizado los niveles de colesterol y de lipoproteínas, como eventos intermedios. En relación al cáncer, los marcadores de efecto precoz utilizados incluyen cambios genéticos en las células, como son la aparición de aberraciones cromosómicas y micronúcleos.

9. Valoración de la exposición

La medida de la exposición es un punto crucial en cualquier estudio epidemiológico. Al hablar de exposición, sin embargo, no nos referimos sólo al posible contacto con un agente de interés, sino también a todos aquellos factores relacionados con el sujeto que pueden influir en la relación entre el agente y la enfermedad. Una adecuada recogida de información sobre hábitos de vida, indicadores de situación socioeconómica, problemas de salud u otras características que varíen la susceptibilidad personal a los agentes de interés y puedan modificar el efecto que se desea

analizar, es casi tan importante como la misma valoración de la exposición al factor que motiva la realización del estudio.

En general, los estudios de cohortes son un diseño muy atractivo para estimar la exposición ya que garantizan que la medición se realiza antes de que se produzca la enfermedad. Así se cumple uno de los principios básicos en causalidad: “la causa precede al efecto”. Esta secuencia temporal tiene otras ventajas añadidas:

1. La medida de la exposición no está condicionada por la enfermedad ni por los tratamientos que se puedan aplicar para ella. Este aspecto es muy importante si se usan marcadores biológicos para estimar la exposición.
2. Las variables de confusión (por ejemplo, aspectos relacionados con la dieta, ejercicio o hábitos de vida) se miden también antes de la aparición de la enfermedad. Así se evitan tanto el sesgo de recuerdo como el posible efecto que el diagnóstico de la enfermedad puede tener sobre las respuestas y conductas de los sujetos. Como contrapartida, sin embargo, hay que señalar que lo habitual es que la información sobre variables asociadas se recoja con menos detalle que en un estudio de casos-controles, en el que se puede focalizar más la atención hacia los aspectos relacionados con el evento de interés. Existe, además, otro problema derivado del momento de recogida de los datos: no podremos tener en cuenta nuevos factores de riesgo que se identifiquen posteriormente, porque no dispondremos de información sobre los mismos. La modificación del instrumento de medida (e.j cambios en el cuestionario) puede permitir tener dicha información a partir de un momento dado. Por ejemplo, el “Nurses’ Health Study” incorporó un cuestionario dietético a partir del cuarto año de seguimiento.
3. Aunque muchas veces es difícil conocer la fecha de la exposición inicial, el diseño de la cohorte permite disponer al menos de una estimación de la exposición en un momento concreto para todos sus miembros. Si el seguimiento de la cohorte es largo, dispondremos de un rango muy amplio de tiempos diferentes entre el momento de la medición y el momento de aparición del proceso de interés. Esto puede permitir investigar hipótesis relacionadas con diferentes periodos de latencia entre exposición y enfermedad.

10. La estimación de la exposición

Un estudio de cohortes es esencialmente un estudio analítico; la estimación de la exposición, por tanto, se realiza con un fin: estudiar la asociación entre aquella

y unos efectos determinados. Esto hace que el investigador tenga que plantearse a priori la relación entre ambas y cuáles son los aspectos a tener en cuenta para representar la “dosis” de exposición relevante en el contexto de esa enfermedad. Para entender a que nos referimos vamos a usar como ejemplo el estudio del efecto a largo plazo de una dosis excesiva de radiaciones ionizantes ocurrida accidentalmente en trabajadores de una central nuclear. La estrategia de medida será diferente en función de las asunciones previas, por ejemplo: a) si presumo que sólo se producirán efectos dañinos en caso de que la radiación haya superado un determinado dintel, b) si lo que presupongo es que todos los trabajadores pueden sufrir consecuencias adversas por este accidente y que a mayor exposición el riesgo de alteraciones es mayor, o c) si considero que el efecto de esta exposición va a estar determinado también por la exposición a radiación que el sujeto acumule a lo largo de su vida. Si precisamente lo que se pretende es caracterizar esta relación, es necesario tener presentes todas estas posibilidades y recoger la información necesaria.

Tras haber pensado cuál es la mejor manera de aproximarse a esa teórica medida de “dosis” ideal, el investigador tiene que identificar qué exposiciones incluyen el agente de interés, definir los ítems que necesita recoger para obtener información sobre la duración, frecuencia e intensidad de la exposición en el periodo específico de riesgo que se ha determinado, y desarrollar el algoritmo que combina dichas variables para llegar al constructo de dosis elegido.

Aunque en este capítulo vamos a tratar dosis y exposición prácticamente como si ambos términos fuesen sinónimos, hay que señalar que muchos autores diferencian claramente entre exposición, referida a medidas tomadas en el ambiente, es decir fuera del sujeto, y dosis, que se mide bien en tejidos humanos o en el punto de contacto entre sujeto y medio. Ott, en la Enciclopedia de Salud y Seguridad en el Trabajo de la OIT, define la exposición como la concentración de un agente en la frontera entre individuo y medioambiente, y la dosis como el depósito o absorción de un agente por unidad de tiempo, y se estima combinando datos de mediciones ambientales con hipótesis tipificadas o con medidas directas. La diferencia entre las dos depende de características fisiológicas, de los patrones de actividad de los sujetos y de las variaciones en la concentración de los agentes en el tiempo y el espacio. En general se considera que las medidas externas son más útiles que las internas cuando las exposiciones están muy extendidas y las diferencias entre grupos son más importantes que las interindividuales.

10.1. ¿Cómo medimos la exposición? Exposición y tiempo:

En el estudio de cohortes más sencillo, la exposición sería una condición permanente y fácilmente identificable, por lo que asignar individuos al grupo de ex-

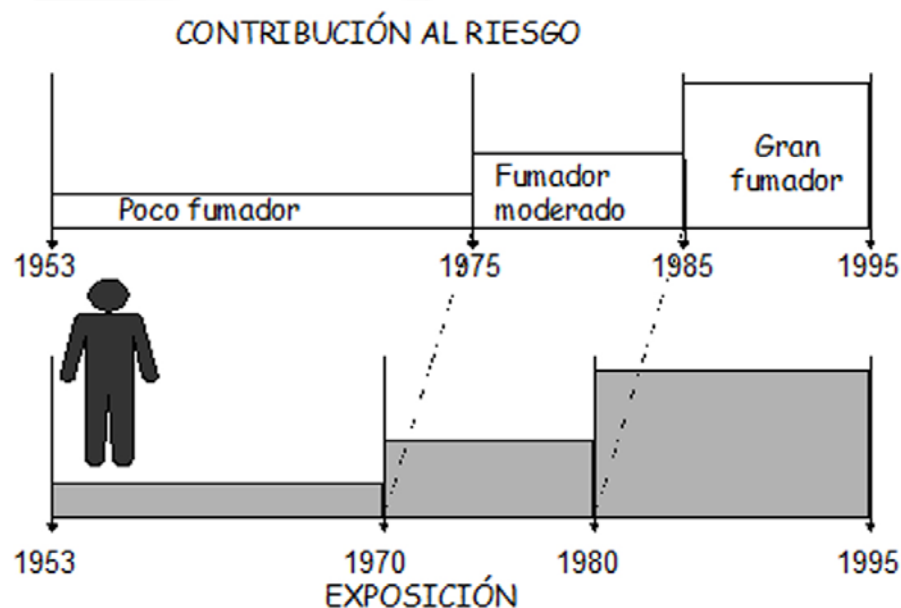
puestos sería prácticamente una tarea rutinaria. Sin embargo, en la realidad las exposiciones de interés rara vez son constantes y a menudo son difíciles de medir. Otros factores de riesgo presentan también este problema: existen características que no cambian en el tiempo, como el sexo, pero otras pueden variar durante el seguimiento, como por ejemplo la ocupación o la residencia habitual.

El primer aspecto a definir es el periodo de exposición que consideramos relevante: ¿una medida de la exposición actual es suficiente o se necesita una medida pasada? Casi todas las medidas de laboratorio (“objetivas”) son actuales, aunque algunas, como las derivadas de rasgos heredados, son fijas. Sin embargo, a menudo es necesario tener en cuenta la posible existencia de periodos de inducción y de latencia, es decir, los periodos de tiempo que se consideran necesarios para que el agente pueda producir el daño (inducción) y para que éste sea detectable (latencia). En la figura 1 se muestra un esquema de la posible relación entre un tóxico y una enfermedad. Si estamos estudiando, por ejemplo, tabaquismo y cáncer de pulmón en mujeres, debemos suponer que ha de pasar algún tiempo desde que la mujer empezó a fumar para que el tabaco pueda generar daño y se origine el cáncer. Al principio, además, el tumor que empieza a desarrollarse en el pulmón es prácticamente indetectable, y ha de pasar un tiempo hasta que produzca síntomas y sea diagnosticado.

Otra pregunta importante es durante que periodo se recogerá la medida de la exposición: ¿es suficiente con una única medida, necesitamos medidas repetidas para aumentar la precisión o es preferible actualizar la información con medidas repetidas del marcador en diferentes momentos desde la medida inicial? De nuevo es fundamental el esquema teórico del investigador para determinar cuál es la estrategia adecuada. En muchos análisis epidemiológicos se usa una medida resumen, asumiendo implícitamente que es éste el único aspecto de la exposición que determina el riesgo actual. Las medidas resumen más habituales son el nivel de exposición en el momento de la medida, la exposición media o mediana si existen medidas repetidas, la exposición máxima registrada o la exposición acumulada, es decir la suma de exposición en cada categoría por el tiempo que el individuo pasa en la misma. El seguimiento en el tiempo de los individuos, sin embargo, conlleva el problema añadido de clasificar la “experiencia” de un individuo en diferentes categorías de exposición en distintos momentos del tiempo. Por ejemplo: en la figura 2 podemos ver cómo se podría categorizar la exposición acumulada al tabaco de un sujeto, que podría pertenecer a la cohorte de médicos británicos previamente mencionada, y que fuma sin interrupción entre 1953 y 1995. Inicialmente el sujeto fuma entre uno y 5 cigarrillos al día, y hasta 1970 no acumula las 1000 cajetillas-vida que hemos definido como límite para ser fumador moderado. Si, además, tenemos en cuenta un posible periodo de latencia de 5 años, consideraríamos que

las personas-año que aporta este sujeto entre 1953 y 1975 corresponden al estrato de poco fumador. Nuestro individuo incrementa algo su consumo de cigarros y en solo 10 años alcanza las 2000 cajetillas-vida que determinan el paso a gran fumador. Las personas-año correspondientes al periodo 1975-85 corresponden a fumador moderado y, a partir de aquí, a gran fumador hasta que el sujeto muera, se pierda o se abandone el seguimiento.

Figura 2. Categorización de la exposición acumulada al tabaco.



10.2. Herramientas para estimar la exposición

La calidad en la medida de la exposición determinará en muchos casos la validez de los resultados del estudio. Por este motivo deben usarse instrumentos de medida correctamente validados, y es conveniente probarlos en una pequeña muestra, revisarlos y probarlos de nuevo en un estudio piloto. En la tabla 3 se recogen las posibles fuentes de error que hay que tener presentes en la estimación de la exposición. La precisión y el grado de detalle con el que se recoge la información es otro aspecto importante. En general hay que intentar, en la medida de lo posible, cuantificar. La información cuantitativa permite estudiar la relación dosis-respuesta y la dosis de exposición mínima necesaria. Es necesario recoger además toda la información sobre las condiciones en las que se produce la exposición si se cree que pueden modificar la misma.

Tabla 3. Posibles fuentes de error en la medida de la exposición

• Error en el instrumento de medida o en su protocolo - Falta de exhaustividad en cubrir todas las fuentes de exposición - Inclusión de exposiciones que no tienen el agente activo - Periodo evaluado distinto del periodo relevante - Preguntas ambiguas o mal entendidas
• Errores en el protocolo de uso del instrumento - Poca especificación de las instrucciones - Imprevisión sobre la forma de manejar situaciones y respuestas no previstas
• Cambios en la precisión a lo largo del tiempo - Cambio del instrumento de medida - Cansancio del entrevistador - Ausencia de estandarización del procedimiento - Degradación de muestras biológicas
• Errores en el procesado de los datos - Errores en la grabación - Errores en la conversión de la exposición al agente activo - Variable indicadora o resumen de la exposición inadecuada

Teniendo presentes todos estos aspectos, el investigador debe elegir un instrumento de medida atendiendo a dos puntos básicos: el factor a estudiar y la factibilidad de su uso en el estudio concreto que se plantea. Entre las formas más habituales de medir la exposición podríamos citar las siguientes:

1. Entrevistas, cuestionarios o, con menor frecuencia, diarios estructurados. Son fundamentales sobre todo a la hora de disponer información relativa a hábitos o conductas. Permiten tener información cuando no existen registros objetivos, aunque pueden estar sujetos a errores difíciles de controlar. Por ejemplo, las contestaciones a determinadas preguntas pueden depender de la consideración social del comportamiento sobre el que se pregunta.
2. Medidas en el ambiente del sujeto, bien procedentes de registros existentes o bien obtenidas expresamente para el estudio (p. ej, registros de contaminación ambiental en una ciudad).
3. Medidas en el entorno inmediato del individuo (por ejemplo, monitores ambientales en una estancia determinada de una empresa).
4. Dosis individual (por ejemplo, monitorización personal).
5. Medidas de concentración en tejidos humanos o productos metabólicos (biomarcadores).
6. Marcadores de efectos fisiológicos (biomarcadores de dosis efectiva).

Muchas veces, en los estudios se combinan varios instrumentos de medida, como se puede comprobar en las cohortes que se han citado como ejemplo. Cada vez es más habitual incorporar recogida de materiales biológicos (sangre, DNA...), acompañados de una encuesta, que resulta imprescindible para tener información sobre otros factores de riesgo o posibles factores de confusión. En la tabla 4 se resumen las estrategias que se han utilizado para medir la exposición en las cohortes que se están usando como ejemplo.

Tabla 4: Instrumento de medida de la exposición y periodicidad de la misma en algunos de los estudios de cohortes citados como ejemplo

Estudio	Instrumento de medida de exposición	Periodicidad de la medida de la exposición	Algunos aspectos interesantes del estudio
EPIC	Cuestionarios postales o en entrevista Examen físico para medidas antropométricas Muestras biológicas: sangre	Cuestionario reducido a los 3-4 años del reclutamiento	El estudio incluyó una exhaustiva investigación metodológica para garantizar la validez de los cuestionarios y medidas utilizados que ha sido publicada independientemente.
NHS-I	Cuestionarios postales	Cada 2 años la información sobre uso hormonal, estatus menopáusico y tabaco.	Ha ido incorporando cuestionarios sobre nuevos aspectos a lo largo del estudio (dieta en 1980, calidad de vida en 1992)
	Muestras biológicas: Recogida de uñas de los dedos de los pies a los 6 años del inicio (68 000 mujeres)	Cada 4 años información sobre dieta y calidad de vida	Algunas tomas de muestras han sido realizadas por los mismos miembros de la cohorte
	Muestra de sangre domiciliaria a los 13 años del inicio (33 000 mujeres aprox.)	Toma única Prevén repetir la muestra en el año 2000	
MACS	Cuestionario autoadministrado Entrevista personal Examen clínico y de laboratorio	Se repite todo cada 6 meses	
Cohorte sueca	Enlace de registros de Cáncer, Mortalidad y censos de 1970 y 1960		Se puede combinar con matrices de ocupación-exposición expresamente diseñadas para esta cohorte
Factores pronósticos CPNM	Encuesta Muestras biológicas	Medida única al inicio del estudio	

Un punto elemental a tener en cuenta si hablamos de cohortes grandes es el coste en dinero y en esfuerzo. Entre las estrategias posibles para realizar estudios de cohortes con costes reducidos podríamos destacar las siguientes:

- a) Los estudios **basados en enlaces de registros**, como la cohorte de población activa sueca, en los que bases de datos que contienen información sobre exposición se asocian a través de números de identificación personales con registros que contienen la información sobre enfermedad. Es bastante habitual y más económico recurrir a registros preexistentes (listados de ocupaciones, etc...), aunque muchas veces esto hace que se disponga de poca información sobre posibles variables de confusión.
- b) Los estudios que reclutan la cohorte y recogen la información sobre exposición a través de **cuestionarios que se envían por correo**, que han de ser necesariamente breves y concisos. Esta es la estrategia que se ha usado, por ejemplo, en el “Nurses’ Health Study”. En este caso, además, se solicitó a los mismos miembros de la cohorte el envío de muestras biológicas (uñas).
- c) los que **miden la exposición sólo en una muestra de la cohorte**. Usan diseños que permiten reducir costes, minimizando la pérdida de información, como veremos más adelante.

10.3. Algunos instrumentos de medida interesantes:

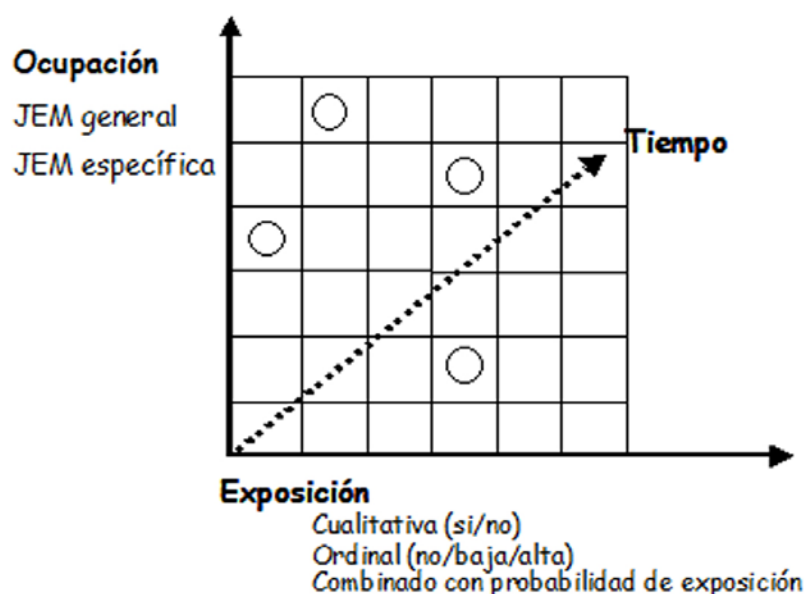
Matrices de Ocupación-exposición

El entorno laboral es una fuente muy rica para el estudio de tóxicos en humanos. En una ocupación, sin embargo, suele haber exposición a múltiples sustancias; y por otra parte existen muchas ocupaciones que tienen exposiciones comunes. En muchos casos, la información relevante es la exposición a la sustancia de interés y no la ocupación desempeñada.

Una matriz de ocupación-exposición (Job-Exposition Matrix: JEM) (figura 3) es una herramienta para convertir las ocupaciones en exposiciones potenciales. En su forma más sencilla, los puestos de trabajo se clasifican en expuestos o no expuestos. La estimación de la exposición en un trabajo concreto puede basarse en medidas directas o en opiniones de expertos basadas en la descripción de los puestos de trabajo, y será más acorde con la exposición real cuanto más específica sea la definición del mismo. Si definimos categorías ocupacionales (p.ej. empleados de tiendas), el error en la estimación será superior al que se tendrá si se estudia cada ocupación de forma independiente en cada sector industrial (p.ej. empleadas de tiendas en gasolineras y empleados de tiendas en un centro comercial). Aún así, no todos los sujetos pertenecientes a una misma combinación ocupación-industria es-

tán expuestos a las sustancias de interés. A veces, la clasificación de la exposición se basa en la proporción estimada de personas en ese empleo que soportan niveles de exposición por encima de un nivel determinado. No obstante, las JEMs no recogen toda la variabilidad que puede existir entre los trabajadores de una misma ocupación, ni la debida a factores personales, por lo que siempre habrá errores de clasificación.

Figura 3. Esquema de una matriz de ocupación-exposición (JEM).



En algunos países, como Suecia, se han elaborado JEMs generales de exposición a químicos o a campos electromagnéticos, que se han aplicado a la cohorte de población activa citada como ejemplo. De esta manera se consigue estudiar la exposición a estos agentes en cohortes muy grandes (nacionales) de una forma barata y eficiente.

Biomarcadores

Un biomarcador de exposición (figura 4) es un estimador de dosis de un determinado agente medido en muestras o sistemas biológicos. Puede ser un compuesto exógeno al individuo, un metabolito de aquél o un producto interactivo exo-endógeno, como las combinaciones de tóxicos con proteínas del sujeto. Resultan especialmente útiles si hay varias vías posibles de exposición (p.ej, ingesta y absorción dérmica), cuando se usan medios de protección o en condiciones de exposición imprevisibles. Otros biomarcadores relevantes para los estudios de cohortes son los marcadores de susceptibilidad genética: diferencias en la velocidad de acetilación

hepática, por ejemplo, hacen que varíe la probabilidad de padecer una enfermedad concreta tras la exposición a un tóxico que se inactiva por esta vía. Modifican por tanto el efecto de la exposición (interacción).

Los biomarcadores presentan un innegable atractivo para los investigadores; sin embargo, como todos los instrumentos de medida, han de ser específicos, fiables y precisos. Sus ventajas y desventajas se resumen en la tabla 5. La adecuada selección del biomarcador debe tener en cuenta: a) con qué periodo de exposición se relaciona el biomarcador seleccionado, b) cómo afectan los factores metabólicos u otros factores personales del huésped a los niveles del biomarcador, y c) las características de la muestra que se precisa (invasiva o no, cantidad de sustrato, posibilidad de degradación en el tiempo, coste, posibilidad de tomas seriadas si es preciso...), ya que determinan la logística del estudio.

Figura 4. Tipos de biomarcadores.

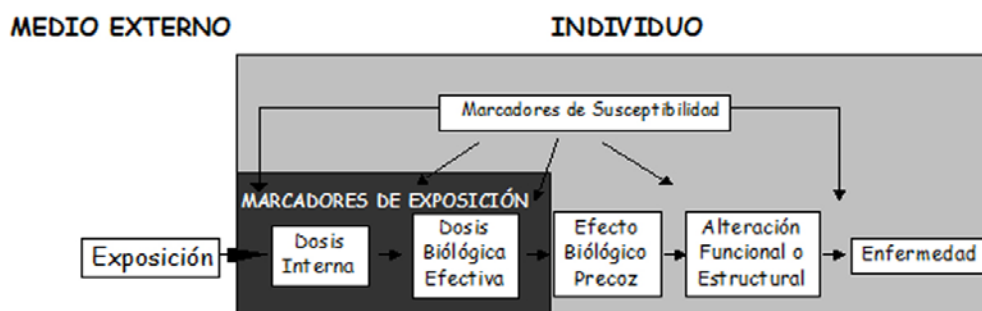


Tabla 5: Biomarcadores en estudios de cohortes: ventajas y desventajas

Ventajas del uso de biomarcadores:

- Puede permitir la integración de varias vías de exposición.
- No hay necesidad de estimar el efecto de las medidas de protección.
- Aunque aparentemente dos sujetos estén expuestos a concentraciones similares de una sustancia, en realidad pueden existir diferencias individuales de "ingesta".
- Por diferencias en exposiciones medioambientales no conocidas.
- Por diferencias en la biotransformación de la sustancia.
- Los biomarcadores pueden dar una idea más real de la exposición concreta de un individuo.
- Si varios agentes tienen un metabolito común pueden estudiar su efecto conjunto.
- Si la variabilidad intra e inter-individuos es menor que la del medio, puede ser una estrategia muy eficiente de medición de la exposición.

Tabla 5: Biomarcadores en estudios de cohortes: ventajas y desventajas (continuación)

Desventajas del uso de biomarcadores
• No siempre hay buena correlación con la exposición ambiental. Es muy importante por tanto validar previamente el marcador. • Cuidado con el periodo al que se refieren: no siempre encontraremos biomarcadores que se correlacionen con el momento de exposición que queremos estudiar. • Los procesos toxicocinéticos y de eliminación modifican las concentraciones de las sustancias de interés y pueden verse alterados por muchas causas (p.ej. exposición simultánea a otro agente no controlado). • Si las diferencias interindividuales por variaciones en la entrada, biotransformación y excreción del tóxico son mayores que las medioambientales, pueden ser mejores las segundas. • Muchas veces no son específicos de un agente concreto: hay que descartar factores de confusión o exposiciones combinadas. • Debemos disponer de niveles base o de referencia para interpretarlos y conocer muy bien sus posibles fuentes de variabilidad. • Puede que no toda la cohorte esté dispuesta a proporcionar muestras, con lo que tendríamos poca potencia. Esto es más evidente para las medidas repetidas, sobre todo si la muestra es invasiva.

11. Otros factores de riesgo: confusión e interacción

La necesidad de obtener información sobre exposición a otros factores viene motivada por dos razones:

1. La naturaleza multicausal de la etiología de la mayor parte de las enfermedades,
2. los factores que determinan la presencia o ausencia de la exposición de interés.

El estudio de cohortes es de carácter observacional, es decir, el investigador estudia la exposición en aquellos grupos de población en los que ésta se produce. Normalmente, la mayor parte de los factores de riesgo no se distribuyen de forma aleatoria en la población, por lo que con frecuencia los determinantes de la enfermedad están asociados, positiva o negativamente, con la exposición bajo estudio. Esta asociación hace necesario identificar y recoger información sobre los posibles factores de confusión para controlar su efecto en el análisis.

Además, en el estudio de la dinámica de la enfermedad es interesante constatar si existen sinergias entre la exposición bajo estudio y la presencia de otros factores de riesgo (presencia de interacciones).

12. Confidencialidad

El establecimiento de una cohorte implica prever un seguimiento para los individuos que la componen. Esto hace imprescindible que exista un identificador para cada sujeto, que permita asignarle la información sobre su exposición o su estatus vital. Sin embargo, como en el resto de los estudios epidemiológicos, asegurar la confidencialidad de la información es un requisito fundamental para cumplir con los principios éticos exigibles a toda investigación. En España, el tratamiento de los datos personales se rige por la ley Orgánica 15/1999 de Protección de Datos de Carácter Personal, que establece las garantías que deben observarse durante la obtención, custodia y cesión, si es el caso, de los mismos. En esta norma, los datos relativos a la salud son protegidos especialmente, con restricciones y controles más estrictos que con otro tipo de informaciones. En el caso de incluir datos sobre enfermedades, es habitual tener que recurrir a las historias clínicas de los sujetos, siendo entonces también de aplicación la ley 41/2002, reguladora de la Autonomía del Paciente y de los Derechos y Obligaciones en Materia de Información y Documentación Sanitaria y la Ley 14/1986, General de Sanidad, que permiten el acceso a la información con fines epidemiológicos o de investigación, aunque explicitan que, como regla general, dicho acceso debe realizarse de forma que quede asegurado el anonimato. En cualquier caso, es responsabilidad del investigador garantizar la confidencialidad de la información que está a su cargo.

13. Estrategias de análisis

El análisis de los estudios de cohortes ha de tener en cuenta la dimensión temporal del estudio. La existencia de pérdidas y la diferente duración del seguimiento de los distintos miembros de la cohorte implican la necesidad de técnicas analíticas especiales que, a la vez, permitan considerar toda la información aportada por los integrantes de la cohorte. Las estrategias de análisis más usadas “desintegran” el tiempo de observación de cada uno de los participantes. La medida de frecuencia más adecuada es la tasa de incidencia. La unidad de análisis ya no es el individuo, sino la unidad temporal de observación. Existen dos estrategias diferentes de análisis en los estudios de cohortes: el análisis supervivencia y el análisis personas-tiempo, pero ambas requieren disponer de:

- Información precisa sobre el inicio y el final del seguimiento de cada participante.

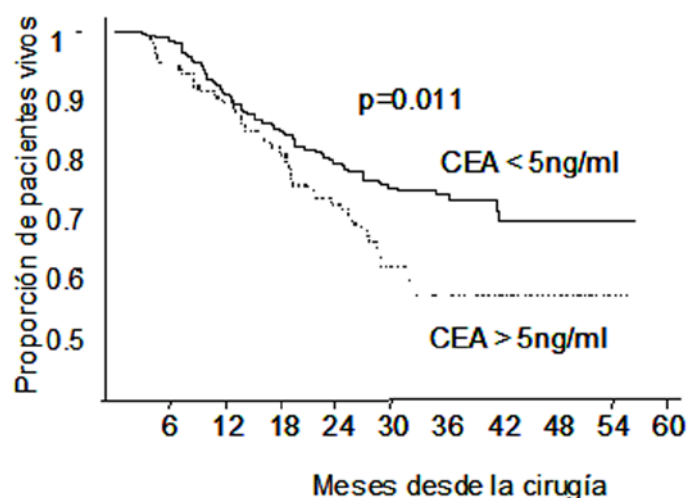
- Definición de una escala temporal adecuada (edad, año calendario, tiempo tras la intervención, etc).
- Estatus del participante en el momento de salida del estudio (desarrollo o no de la enfermedad).

Es importante tener en cuenta, que en el análisis, todos los individuos con información incompleta, es decir, aquellos que no desarrollan el evento de interés, son tratados de la misma manera. Conceptualmente, sin embargo, cabe establecer diferencias entre los participantes que están presentes en el momento del cierre del estudio y aquellos que han abandonado el mismo a lo largo del seguimiento. Las técnicas de análisis consideran que las pérdidas son aleatorias, pero las personas que abandonan el estudio pueden hacerlo por razones o factores relacionados con la enfermedad estudiada o con la exposición de interés. También es importante tener en cuenta la mortalidad competitiva, que impide llegar a diagnosticar la enfermedad en aquellos individuos que mueren por otras causas. Por ello, los estudios de cohortes deben proporcionar información sobre el porcentaje y el tipo de pérdidas observadas a lo largo del seguimiento.

13.1. Las técnicas de análisis de supervivencia

Utilizan la información individual de cada participante de la cohorte. Las técnicas más conocidas son las curvas de supervivencia de Kaplan-Meier, cuya significación estadística es comprobada a través del test de logrank, y los modelos semi-paramétricos de Cox. La popularidad de ambos métodos se basa en la ausencia de asunciones sobre la evolución de la tasa de incidencia en el tiempo. Sin embargo, en ambos casos (logrank y Cox) se asume que la medida del efecto, es decir, la razón de tasas de incidencia (denominada “Hazard Ratio” en los estudios de supervivencia) es constante a lo largo del tiempo (asunción de proporcionalidad). De todas formas, el modelo de Cox permite parametrizar la falta de proporcionalidad dentro del propio modelo. También es posible incluir dentro del modelo exposiciones cambiantes en el tiempo. En la figura 5 se muestra, como ejemplo, la comparación entre las curvas de supervivencia obtenidas en la cohorte de CPNM en enfermos con niveles preoperatorios bajos de antígeno carcinoembrionario (CEA) y en pacientes con niveles elevados. La diferencia entre ambas es estadísticamente significativa.

Figura 5. Comparación de las curvas de supervivencia postoperatoria entre los pacientes del estudio de factores pronósticos del cáncer de pulmón no microcítico según niveles séricos preoperatorios de antígeno carcinoembrionario:



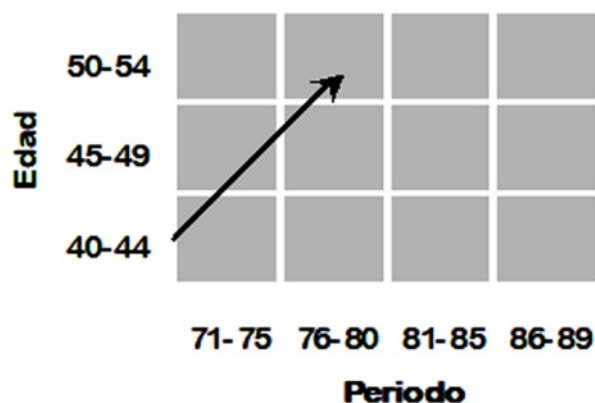
Para el análisis de supervivencia es importante que la unidad temporal esté referida al tiempo biológicamente relevante, que puede o no coincidir con la cronología del estudio. Por ejemplo, en el estudio MACS no tiene mucho sentido considerar el inicio del estudio como origen en el análisis de supervivencia, ya que los individuos seroprevalentes se encuentran en una fase diferente de la enfermedad, mientras que en los seronegativos el momento de la serconversión es el momento relevante.

13.2. Las técnicas de personas-año

Utilizan como unidad de análisis estratos o grupos homogéneos en los que es razonable asumir una misma tasa de incidencia. Esto requiere colapsar la información individual por estratos de interés. Estas técnicas son utilizadas en los grandes estudios poblacionales, como la cohorte de población activa sueca. Las variables que definen los estratos de interés son, lógicamente, la exposición y el resto de factores de riesgo (factores de confusión) a considerar. Algunas de estas variables, como el sexo, pueden ser fijas, mientras que otras (la edad o el periodo de observación, por ejemplo) cambian a lo largo del tiempo. La existencia de estas variables cambiantes en el tiempo hacen necesario “dividir” el tiempo de observación de cada persona para considerar su contribución a cada uno de los estratos (grupos de edad, por ejemplo) por los que va pasando. Si tomamos un sujeto de la cohorte sueca con 43 años en 1971 (ver figura 6) y que es diagnosticado de cáncer de mama en 1978, podemos ver, por ejemplo, que su contribución en personas-año correspondiente al periodo 71-75 debe ser dividida en dos porciones: una parte

corresponde al grupo de edad de 40-44 años y otra parte al grupo de edad 45-49. En la cohorte sueca este proceso fue realizado en cada sujeto de la misma.

Figura 6. Contribución de un individuo a diferentes estratos en el cálculo de las personas año.



En cohortes históricas poblacionales, generalmente los estratos a considerar vienen determinados por las variables para las que se dispone de información. En estos estudios, como ocurre en los estudios descriptivos de grandes colectivos de población, con frecuencia el denominador de las tasas no se calcula de forma exacta, contabilizando la contribución de cada individuo, sino considerando la población a mitad del periodo multiplicada por la duración del periodo.

Las técnicas clásicas de análisis “personas tiempo” incluyen la comparación de tasas, la estandarización directa y las razones estandarizadas de incidencia o mortalidad (denominadas en la literatura RIE o SIR, para la incidencia y RME o SMR para la mortalidad). RIE y RME comparan el número de casos observados con el número esperado teniendo en cuenta la frecuencia de enfermedad en la población utilizada como referencia. La técnica multivariante por excelencia para la modelización de las tasas es la regresión de Poisson, que utiliza la distribución de Poisson para modeliza la variabilidad aleatoria del numerador de las tasas.

14. Potencia estadística

El alto coste de los estudios de cohortes en términos de esfuerzo, tiempo y dinero hace necesario la estimación previa del tamaño muestral requerido en relación al efecto que se pretende observar. En muchas situaciones, sin embargo, es

más práctico proceder al revés. Es decir, considerar en primer lugar el número de participantes que es posible incluir en el estudio, bien sea por restricciones económicas o de otro tipo, y calcular la potencia estadística disponible. Por ejemplo, en el estudio sobre factores pronósticos de CPNM, el volumen de pacientes atendidos en los hospitales participantes indicaba la posibilidad de incluir alrededor de 400 casos durante el reclutamiento. Teniendo en cuenta la proporción de muestras p53 positivas en series previas de pacientes, dicho tamaño muestral suponía, con un periodo de seguimiento mínimo de dos años, la posibilidad de detectar significación estadística (potencia superior al 80 %) para riesgos relativos (hazard ratios) iguales o superiores a 1,50.

Una vez concluido el estudio, el cálculo de la potencia estadística a posteriori es particularmente importante para enjuiciar los resultados negativos, es decir, la ausencia de asociación entre la exposición y la enfermedad. Los métodos estadísticos no permiten nunca demostrar la hipótesis nula (falta de asociación) sino sólo rechazarla o no con un margen de error determinado. Por ello, cuando la incidencia del evento de interés es similar en expuestos y no expuestos, es importante tener en cuenta las siguientes consideraciones:

- ¿Cuál es la potencia estadística real del estudio? ¿Cuál es el mínimo RR que sería posible detectar teniendo en cuenta las características del mismo?
- ¿Existen diferencias reales de exposición en los grupos de comparación?
- ¿El periodo de seguimiento ha sido suficiente? Hay que tener en cuenta el periodo de latencia necesario (ver figura 1).
- ¿Cómo encajan los resultados del estudio en el marco general del conocimiento existente sobre dicha enfermedad?

15. Problemas de interpretación de los estudios de cohortes

15.1. Representatividad de la cohorte seleccionada

Este problema puede estar presente cuando se seleccionan grupos específicos, por su disponibilidad o su mayor tasa de participación, para representar a toda la población (el estudio de los médicos británicos, o de las enfermeras americanas, por ejemplo). La inclusión en los estudios de voluntarios, es decir, personas que aceptan participar, también puede conllevar problemas de representatividad (además de los estudios citados anteriormente, el EPIC y el MACS utilizan voluntarios).

Los voluntarios suelen ser personas con mayor nivel de educación, más preocupados por su salud y con hábitos de vida más saludables. Esto puede implicar una infraestimación de exposiciones nocivas y una menor incidencia de la enfermedad.

15.2. Elección del grupo de comparación

Las personas expuestas y no expuestas pueden diferenciarse respecto a otros factores relevantes para la enfermedad de interés. Cuando estos factores son conocidos, el estudio debe recoger información sobre ellos y obtener estimadores del efecto de la exposición, eliminando el posible efecto de confusión de estas variables. En estudios que utilizan grupos externos de comparación, esta información no suele estar disponible.

15.3. Pérdidas durante el seguimiento

Como ya se ha comentado, las técnicas de análisis consideran todos los individuos que no llegan a desarrollar la enfermedad como un grupo homogéneo, pero las causas de esta falta de información son diversas. El principal problema radica en los sujetos que abandonan el estudio y en aquellos que mueren por otras causas. Las técnicas analíticas asumen que las pérdidas son aleatorias y no están relacionadas con la enfermedad o la exposición de interés. Por ello, es importante comprobar que las personas que permanecen en el estudio y aquellas que lo han abandonado son realmente similares. Cuando las pérdidas no son aleatorias se produce un sesgo de selección a lo largo del tiempo y la cohorte deja de representar a la población de base inicial.

Sesgos introducidos por los errores de medida:

Ya se ha hablado de ellos en el apartado dedicado a la exposición (ver también la tabla 3). Es importante que dichos errores sean de tipo “no diferencial”, es decir, que el error esperado sea similar en los distintos grupos de exposición.

15.4. El problema de las comparaciones múltiples

Este problema se presenta en aquellos estudios de cohortes que valoran múltiples exposiciones o múltiples eventos de interés. Por ejemplo, en la cohorte de población activa sueca, el interés se centra en conocer la incidencia relativa de un determinado tipo de cáncer en cada grupo ocupacional. Esto supone la realización de múltiples contrastes, comparando cada una de las 270 ocupaciones con el resto de la población activa. En estos estudios, la posibilidad de encontrar, por azar, asociaciones estadísticamente significativas aumenta considerablemente. Reciente-

mente se han propuesto nuevos procedimientos para recalcular la significación estadística teniendo en cuenta el número de comparaciones realizadas, minimizando la “tasa de hallazgos falsos” (False Discovery Rate).

16. Casos y controles basados en estudios de cohortes

Se trata de diseños de estudios caso-control que pretenden maximizar la información obtenida a partir de los estudios de cohortes minimizando el coste. Los principales diseños son dos: los casos y controles anidados a una cohorte y los estudios caso-cohorte. Estos diseños son especialmente importantes cuando el estudio incluye determinaciones analíticas costosas o muestras difíciles de obtener, como por ejemplo, biomarcadores de exposición o susceptibilidad.

16.1. Estudios de casos y controles anidados

Son estudios caso-control que incluyen todos los casos observados en la cohorte y uno o más controles emparejados con los casos por tiempo de seguimiento. Este apareamiento implica considerar la información disponible para los controles sólo hasta la fecha de diagnóstico de su caso correspondiente. Además, un control puede llegar a convertirse en caso, a medida que avanza el seguimiento. El análisis de los datos debe respetar el apareamiento y comparar cada caso con su grupo de controles.

16.2. Estudios caso-cohorte

Bajo este diseño se seleccionan como controles, de forma aleatoria, una fracción de todos los componentes de la cohorte original. Este tipo de diseño supone la consideración de la información proporcionada por la subcohorte seleccionada más la de todos los casos ocurridos fuera de dicho grupo. En este tipo de diseño, los métodos de análisis deben tener en cuenta el carácter diferencial de los casos incluidos y no incluidos en la subcohorte seleccionada.

17. Estudios de cohortes y los ensayos clínicos

La diferencia entre estos dos diseños es la asignación aleatoria de la exposición, propia de los ensayos clínicos. Dicha asignación aleatoria, o aleatorización, garantiza la comparabilidad de los grupos de exposición siempre que el tamaño muestral sea adecuado.

Por razones éticas obvias, los ensayos clínicos sólo se utilizan para valorar exposiciones o intervenciones potencialmente beneficiosas. En estas circunstancias, los ensayos clínicos son preferibles a los estudios de cohortes. La aleatorización sitúa a los ensayos clínicos en el marco de investigación propio de los estudios experimentales, donde el investigador puede controlar las condiciones en las que se produce el experimento.

Sin embargo, cuando se investiga la utilidad de nuevos tratamientos, los criterios de inclusión utilizados en los ensayos clínicos suelen implicar una gran selección de los pacientes. Con frecuencia, los grupos más vulnerables (pacientes con otras patologías subyacentes, edades avanzadas, grupos sociales desfavorecidos, etc) no forman parte de estos estudios. Además, las condiciones de aplicación del tratamiento, el seguimiento y el manejo de los casos dentro de un ensayo clínico pueden no ser representativos del contexto clínico general en el que se diagnostica y trata a los pacientes. Estas diferencias explican los resultados paradójicos encontrados, por ejemplo, cuando se evaluó la utilidad de los primeros agentes antirretrovirales. Mientras los ensayos clínicos mostraban un efecto beneficioso, la mortalidad de los casos con SIDA de estudios de cohortes como el MACS no mostraban diferencias sustanciales entre los pacientes tratados y no tratados. Hubo que esperar a la llegada de la nueva terapia potente antirretroviral para que el impacto de dichos tratamientos fuese realmente visible en las cohortes de pacientes con SIDA.

Teniendo en cuenta lo anterior, es interesante matizar que los ensayos clínicos determinan la eficacia teórica de un determinado tratamiento. Como contrapartida, los estudios de cohortes representativas de pacientes juegan un papel importante en el estudio de la efectividad, es decir, el efecto real de dicho tratamiento aplicado en las condiciones normales de la práctica clínica.

En otro tipo de contexto, la aleatorización resulta inviable al existir grupos naturalmente expuestos y grupos no expuestos en la población. Los estudios de cohortes son, en este caso, la mejor aproximación de que dispone el epidemiólogo para acercarse a la calidad de la información generada por los ensayos clínicos.

18. Conclusiones

Como se ha comentado en este capítulo, los estudios de cohortes suponen una herramienta esencial en la investigación epidemiológica. En ellos, la relación temporal causa-efecto está asegurada, siendo posible además investigar e incorporar en el análisis los cambios que se producen en la cohorte a lo largo del tiempo.

El diseño y desarrollo de un estudio de cohortes es generalmente un reto que conlleva muchas dificultades. Es necesario valorar la exposición y los cambios que se producen a lo largo del tiempo, retener a los participantes en la cohorte y mantener un sistema de seguimiento que permita determinar la incidencia de la enfermedad o enfermedades estudiadas.

Los estudios de cohortes han demostrado ser útiles no sólo a la hora de investigar los factores etiológicos más importantes de muchas enfermedades, sino también para caracterizar la evolución de epidemias como el SIDA, predecir el riesgo de enfermar a nivel individual (por ejemplo, las fórmulas que valoran el riesgo de enfermedad cardiovascular proceden de la cohorte Framingham y de estudios similares), investigar el efecto sinérgico de dos o más exposiciones, determinar el periodo de latencia, investigar la validez de los marcadores precoces de enfermedad, etc. Finalmente los estudios de cohortes complementan a los ensayos clínicos a la hora de valorar la utilidad de los nuevos agentes terapéuticos. Los estudios de cohortes han servido además de estímulo en el planteamiento de nuevas hipótesis y el desarrollo de nuevos métodos de análisis. La riqueza de la información aportada por estos estudios, la capacidad que han demostrado para incorporar nuevas hipótesis y ayudar a clarificar el papel desempeñado por diferentes factores de riesgo en la génesis de enfermedades complejas sitúan a estos estudios como marco de referencia en la investigación epidemiológica.

19. Referencias

1. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B* 1995; 57:289-300.
2. Breslow NE, Day NE. *Statistical Methods in Cancer Research Volume II : The design and analysis of cohort studies*. IARC Scientific Publications No. 82. International Agency for Research on Cancer. Lyon, 1987.

3. Colditz GA, Manson JE, Hankinson SE: The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J Women's Health* 1997; 6: 49-62.
4. Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. *Br Med J* 1964; 1: 1399-1410.
5. Guallar EM. Alternativas al diseño tradicional de los estudios de cohortes: estudios de casos y controles híbridos y estudios de cohortes y casos. *Rev Salud Pública* 1991; 2: 151-165.
6. Hunter, D.J. Methodological issues in the use of biological markers in cancer epidemiology: cohort studies. En Toniolo P. y cols.. Ed. *Application of Biomarkers in Cancer Epidemiology*. IARC Scientific Publications N.º142. International Agency for Research on Cancer. Lyon, 1997.
7. Kauppinen T, Toikkanen J y Pukkala E. From cross-tabulations to multi-purpose exposure information systems: a new job-exposure matrix. *Am J Ind Med* 1998; 33:409-417
8. Lyles RH, Munoz A, Yamashita TE, Bazmi H, Detels R, Rinaldo CR, Margolick JB, Phair JP, Mellors JW. Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS Cohort Study. *J Infect Dis.* 2000;181:872-80.
9. Muñoz A, Hoover DR. Use of cohort studies for evaluating AIDS therapies. En: Finkelstein DM, Schoenfeld DA (eds). *AIDS Clinical Trials*. Wiley and Sons. NY 1995. pp: 423-446.
10. Nieto-García FJ. Los Estudios de cohorte. En: Martínez-Navarro F, Antó JM, Castellanos PL, Gili M, Marset P, Navarro V (eds). *Salud Pública*. McGraw-Hill y Interamericana. Madrid, 1999.
11. Organización Internacional del Trabajo. Enciclopedia de Salud y Seguridad en el Trabajo. 3.ª Ed. Ministerio de Trabajo y Asuntos Sociales 2001. <http://www.mtas.es/insht/EncOIT/Index.htm>
12. Pollán M, Gustavsson P. Cancer and Occupation in Sweden: 1971-1989. EPC-Rapport 1999:1. The National Board of Health and Welfare, Sweden, 1999. Disponible en: <http://www.sos.se/fulltext/9918-001.pdf>
13. Pollán M. Ocupación, exposición laboral a radiaciones electromagnéticas y cáncer de mama. Instituto de Salud Carlos III. Madrid, 2001. Disponible en: <http://193.146.50.130/cancer/cem-cancer.pdf>
14. Pollán M, Varela G, Torres A, de la Torre M, Ludeña MD, Ortega MD, Pac J, Freixenet J, Gómez G, Sebastián F, Díez M, Arrabal R, Canalís E,

- García-Tirado J, Arnedillo A, Rivas JJ, Minguella J, Gómez A, García M, Aragonés N, Pérez-Gómez B, López-Abente G, González-Sarmiento R, Rojas JM. Clinical value of p53, c-erbB-2, CEA and CA125 regarding relapse, metastasis and death in resectable non-small cell lung cancer. *Int J Cancer* (in press).
15. Riboli E, Kaaks R. The EPIC Project: rationale and study design. *European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol* 1997 26: S6-S14.
 16. Rothman KJ, Greenland S. Cohort Studies. En: *Modern Epidemiology*. Second Edition. Lippincott-Raven. Philadelphia, 1998. pp 79-91.
 17. Samet JM, Muñoz A (eds). Cohort Studies. *Epidemiologic Reviews* 1998; 20, 1.
 18. Schottenfeld D, Fraumeni FJ, Jr. *Cancer Epidemiology and Prevention*. Second Edition. Oxford University Press. Oxford, 1996.
 19. Tarwater PM, Mellors J, Gore ME, Margolick JB, Phair J, Detels R, Muñoz A. Methods to assess population effectiveness of therapies in human immunodeficiency virus incident and prevalent cohorts. *Am J Epidemiol*. 2001;154: 675-81.

Capítulo 11.

Estudios de casos y controles

LAURA P. DEL NIDO VARO, MARINA POLLÁN SANTAMARÍA
Y BEATRIZ PÉREZ GÓMEZ

1. ¿Qué es un estudio de casos y controles?
2. ¿Cuándo utilizar un estudio de casos y controles?
3. Los puntos clave en los estudios de casos y controles
4. La selección de los casos
5. La selección de controles
6. La medida de la exposición
7. Medidas de asociación
8. Confusión y sesgos en estudios de casos y controles
9. Emparejamiento
10. Ejemplos de estudios de casos y controles
11. Referencias

Hay muchas circunstancias en las que para evaluar la relación entre un determinado factor y una enfermedad no es factible o éticamente aceptable utilizar un diseño experimental (por ejemplo, no se puede hacer un ensayo clínico para ver si una sustancia es o no cancerígena), y en las que diseñar un estudio de cohortes puede ser difícil y requerir mucho tiempo de seguimiento (por ejemplo, para enfermedades infrecuentes o si hay necesidad de resultados rápidos). Un diseño alternativo para estos casos son los estudios de casos y controles. Son una herramienta muy útil en epidemiología, ya que su elegante diseño permite, en muchas situaciones, extraer conclusiones válidas haciendo un uso eficiente de los recursos disponibles. En estos estudios identificamos enfermos (casos) y no enfermos (controles), y evaluamos si han estado o no expuestos a nuestro factor de interés. Esta aparente facilidad de ejecución, sin embargo, lleva como contrapartida mayor riesgo de incorporar sesgos que los estudios de cohortes. Esta percepción ha llevado a veces a algunos investigadores a denostar un tipo de estudio que, si se usa adecuadamente, no es en modo alguno inferior a otros.

1. ¿Qué es un estudio de casos y controles?

La característica definitoria de los estudios de casos y controles es que la selección de los sujetos se hace en función de su condición de enfermos o no enfermos, entendiendo “enfermo” de forma genérica, incluyendo todo tipo de desenlaces que se quiera estudiar. Son estudios observacionales, ya que el investigador no asigna la exposición —sólo recoge información de lo que ha ocurrido—, y pueden ser analíticos o descriptivos (en función de si su objetivo es comprobar hipótesis o describir asociaciones).

Para poner este tipo de diseño epidemiológico en perspectiva, vamos a plantear, por ejemplo, una investigación que busque saber si una exposición (por ejemplo, desayunar café con leche) es causa de una enfermedad o problema (por ejemplo, hipertensión arterial o cáncer de estómago). Idealmente seleccionaríamos un grupo de personas sin enfermedad, en este caso sin cáncer, las repartiríamos al azar en dos grupos y daríamos café a uno de ellos y al otro no, siguiéndolas en el tiempo y comparando cuántos desarrollan cáncer (o hipertensión arterial, en el primer ejemplo) en cada grupo (*estudio experimental*). Nuestra siguiente alternativa, si esto no es posible, es identificar personas sin cáncer, clasificarlas en dos grupos —las que beben café habitualmente y las que no lo beben— y seguirlas en

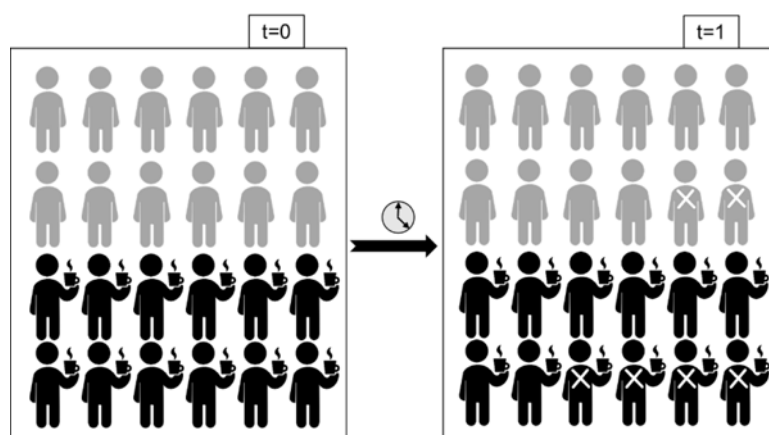
el tiempo, de nuevo comparando cuántas personas de cada grupo desarrollan cáncer de estómago (*estudio de cohortes*). El tercer diseño posible sería empezar por el final, identificando primero a las personas que se van diagnosticando de cáncer de estómago en una población; seleccionando, además, una muestra de esta misma población y, finalmente, buscando luego la información sobre si tomaban o no café las personas de ambos grupos (*estudio de casos y controles*).

Como se ve, los estudios de casos-controles conllevan un **paso de muestreo extra** si se comparan con los otros dos diseños: los estudios de cohortes o experimentales estudian grupos completos de personas susceptibles de enfermar, mientras que en el estudio de casos y controles, nuestros controles son una muestra de la población que sirve para representar cuál sería la frecuencia de la exposición en la población de interés (la naturaleza de este segundo muestreo se verá con más claridad en el siguiente apartado). Por eso, uno de los puntos clave en este tipo de estudios es la selección de los controles. El segundo punto clave es cómo medir la exposición, ya que ha ocurrido en el pasado. Este punto lo abordaremos más adelante.

Como hemos visto, el diseño de casos y controles es, en el fondo, un estudio de cohortes en el que se identifica a los casos y se selecciona como controles a una muestra de los miembros de la cohorte, susceptibles de enfermar, que no están enfermos. Algunos lo definen como un estudio de cohortes al revés (“trohoc”), y otros consideran que, en el fondo, es un estudio de cohortes mucho más eficiente.

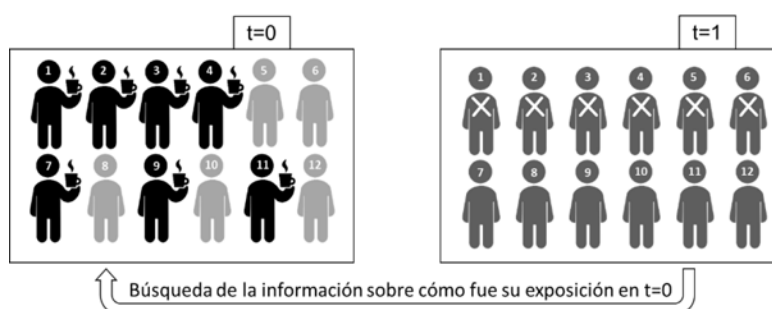
Siguiendo con el ejemplo del café previo, imaginemos una población de personas que, en nuestro caso, contiene 12 sujetos bebedores de café (expuestos) y 12 no expuestos en un momento dado ($t=0$) y que podrían contraer una enfermedad X. Si nos interesara calcular incidencias acumuladas en esta población, observaríamos a todos los sujetos siguiéndolos hasta el final del estudio ($t=1$), diferenciando en ese momento a los que siguen sanos y a los que han enfermado (marcados con una X en la figura 1).

Figura 1. Esquema de un estudio de cohortes.



Para el diseño de casos y controles, sin embargo, partimos de $t=1$, y seleccionamos a los sujetos enfermos (de nuevo, los casos, marcados con X) y solo a unos cuantos de los sujetos sanos (controles), dejando fuera a gran parte de los controles que componían la cohorte originalmente en $t=0$. Posteriormente, buscaremos la información que nos permita saber cuál fue su exposición en el pasado ($t=0$) (Figura 2).

Figura 2. Esquema de un estudio de casos y controles.



La primera pregunta que surge aquí es ¿cómo elegir qué sujetos van a actuar de control? ¿Qué características deberían tener? En otras palabras, ¿cuál es el papel de los controles en un estudio de casos y controles? La finalidad de los controles es **representar la exposición en la población en la que se originaron los casos** (esto aplica también a la exposición a las posibles variables de confusión, de lo cual hablaremos más adelante). Por este motivo, es especialmente relevante prestar una especial atención a la estrategia de reclutamiento de los controles, porque puede comprometer la validez del propio estudio.

Para entender la relevancia de este concepto, y siguiendo nuestro ejemplo, partimos de la cohorte de la figura 1, y nos encontramos en $t=1$. Vamos a imaginar un investigador A y un investigador B (Figura 3). Ambos seleccionan a los mismos casos, pero escogen diferentes controles. El investigador A invita a participar a los controles de la población que salen de una cafetería. El investigador B busca los teléfonos de los miembros de la cohorte y contacta a una muestra aleatoria de ellos. Luego ambos investigadores buscan la información tanto para casos como para controles para clasificarlos como bebedores o no bebedores de café.

Figura 3. Ejemplo: diferente selección de controles en casos y controles.

Como era de esperar, la proporción de bebedores de café entre los controles del investigador A es de un 100 %, y no representa bien a la exposición de la población de origen, que era de un 50 %. Es fácil pensar que el modelo de reclutamiento que ha tenido este investigador ha favorecido que los controles que beben café (expuestos) tengan más probabilidad de entrar en el estudio, dando lugar a un sesgo de selección (también profundizaremos en esto más adelante). Por esto, es muy importante que garanticemos que **la probabilidad de ser reclutado como control sea independiente del estatus o la probabilidad de exposición**. Intuitivamente, parece claro que la estrategia para seleccionar la muestra del investigador B responde mejor a este requisito, ya que los controles tienen una frecuencia de exposición similar a que la que se observaba en la población completa en $t=0$. Los resultados del investigador B serán más parecidos a los que obtendríamos en el estudio de cohortes.

2. ¿Cuándo utilizar un estudio de casos y controles?

No hay una respuesta única. A continuación, están algunos aspectos a tener en cuenta a la hora de determinar si es conveniente o útil tratar de responder a una pregunta de investigación mediante un diseño de casos y controles.

2.1. Ventajas

- Permiten estudiar la relación entre una enfermedad y una exposición (razones de odds).

- Eficientes en enfermedades raras (nos centramos en los individuos informativos).
- Eficiente cuando el periodo de inducción entre exposición y enfermedad o el periodo de latencia son prolongados (es más rápido porque los casos ya se han producido).
- No es susceptible a pérdidas de seguimiento.
- Bajo coste en comparación con otros diseños.
- Permiten estudiar diferentes causas de una misma enfermedad.

2.2. Desventajas

- En muchos casos, no permiten calcular estimadores de incidencia
- Son poco eficientes si la exposición a estudio es rara.
- Generalmente sólo permiten estudiar una enfermedad.
- Son más susceptibles a diferentes tipos de sesgos que los estudios de cohortes (sesgo de selección, sesgo de información, sesgo de supervivencia, sesgo de causalidad inversa o protopático...).

3. Los puntos clave en los estudios de casos y controles

En el diseño de los estudios de casos y controles hay tres aspectos fundamentales, que determinan la calidad del estudio y, por tanto, la validez de las conclusiones del mismo:

- a) La **selección de los casos**, que muchas veces es más compleja de lo que intuitivamente se piensa.
- b) La **selección de los controles**, uno de los aspectos más delicados de este tipo de estudios, y
- c) La **medida de la exposición**, que entraña dificultades propias, ya que busca saber lo que ocurrió en el pasado con la mayor fiabilidad posible, y que debería hacerse de la misma forma tanto en los casos como en los controles.

4. La selección de los casos

Si definimos como caso a toda persona que cumpla un determinado criterio diagnóstico perteneciente a una población definida en un periodo de tiempo definido, idealmente querríamos incluir en nuestro estudio a **todos los casos que ocurren en esa población durante ese periodo de tiempo**. Si esto no es posible, nuestra segunda opción sería hacer un **muestreo entre todos los potenciales casos**, aunque muchas veces es difícil identificar cuál es la población de origen de los casos. Este problema no lo tienen los llamados *diseños híbridos (caso-cohorte y casos-control anidado)*, que son diseños de estudio que se aplican en cohortes bien definidas, en donde se usan los casos que ocurren dentro de la cohorte y se selecciona un reducido número de participantes de la cohorte como control. En cualquier caso, es necesario recordar que la probabilidad de selección tiene que ser independiente tanto de la exposición como de las variables de interés que se deseen incorporar en el estudio.

Siempre es **preferible seleccionar casos incidentes a prevalentes** (1) si queremos ser capaces de estudiar el espectro completo de la enfermedad. Imaginemos estudiamos una enfermedad que tiene casos graves, que mueren poco tiempo después del diagnóstico, y casos leves que viven más tiempo. Si yo selecciono los casos que llevan un determinado tiempo en seguimiento en el hospital, mi estudio estará excluyendo a los casos graves, porque habrán fallecido. Es posible, además, que los casos prevalentes hayan cambiado su exposición tras enterarse del diagnóstico de enfermedad, (por ejemplo, un sujeto que deja de fumar al conocer su diagnóstico de cáncer de pulmón). Lo ideal es reclutar todos los casos que aparezcan, y lo más cerca posible del momento diagnóstico; de hecho, en aquellos diseños que lo permiten, se busca incluso reclutar a los participantes antes de que ellos mismos conozcan su diagnóstico. De esta forma, la enfermedad y su tratamiento no influyen en el comportamiento ni en la forma en la que se contestan los cuestionarios. No obstante, es frecuente el uso de casos prevalentes en el estudio de malformaciones congénitas o en enfermedades crónicas con supervivencia larga en las que resulta especialmente difícil determinar el momento de inicio de la enfermedad (i.e. leucemia linfática crónica, diabetes mellitus, esclerosis múltiple).

5. La selección de controles

5.1. Selección

La situación más habitual a la hora de llevar a cabo un estudio de casos y controles es partir de una definición de casos y diseñar una estrategia para obtener los controles más adecuados para ellos. Aun a riesgo de ser reiterativos, recordemos que la finalidad de seleccionar controles es poder estimar el nivel de exposición en la población origen de los casos. En este punto, el investigador debe hacer un esfuerzo para imaginar cual es la cohorte original que ha dado lugar a los casos y tratar de obtener controles que sean representativos de dicha cohorte.

Siguiendo a Vandenbroucke y Pearce, en un artículo muy recomendable que se incluye en la bibliografía de este capítulo (2) podemos diferenciar dos situaciones diferentes: una, cuando el reclutamiento de controles se hace en una cohorte abierta, y otra, cuando los controles los obtenemos de una cohorte cerrada bien definida.

5.1.1. *Partiendo de una cohorte abierta dinámica en estado estacionario*

La mayoría de las veces, la referencia en los estudios de casos y controles es una cohorte abierta, en una población dinámica, donde hay tanto entradas (nacimiento, migraciones...) como salidas (defunción, movimientos migratorios, fin del periodo de susceptibilidad...). Si la población tiene similar número de entradas y salidas, y la exposición es estable y no cambia en el tiempo, nos encontramos ante una cohorte abierta en estado estacionario. En esta situación los controles son una buena representación de la exposición en la población independientemente del momento. En este escenario, la odds ratio (*OR*) es un buen estimador de la razón de tasas de incidencia (*RT*). Si la exposición se modifica en el tiempo, la selección tiene que adaptarse para tener esto en cuenta, por ejemplo, reclutando a mitad del periodo o emparejando casos y controles por fecha de inclusión en el estudio.

5.1.2. *Partiendo de una cohorte cerrada*

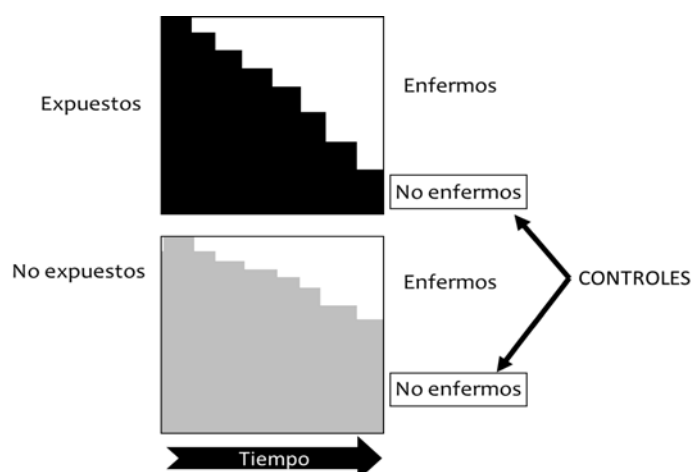
Una cohorte cerrada sería aquella en la que la pertenencia se determina en función de un evento o característica, por ejemplo, una exposición que se produce en un momento dado, a partir de la cual da comienzo un periodo de seguimiento. Como esta característica no cambia con el tiempo, la cohorte es fija. En ella seleccionamos los casos que se producen y una muestra de controles. En todos los casos, podemos calcular la *OR*, que será un buen estimador de la *OR* poblacional, pero nuestro objetivo, en realidad, es usar esta medida como un estimador de ries-

go. Hay tres maneras posibles para reclutar a los controles, que estiman medidas de asociación algo diferentes: (4,5)

- a) Al final del periodo de estudio, muestreando entre aquellos sujetos que para entonces no han desarrollado la enfermedad, es lo que son los **estudios de casos y controles clásicos, o de muestreo acumulativo (Figura 4)**.

Esta estrategia tiene un problema. Imaginemos una cohorte con un 50 % de personas expuestas a un factor de riesgo muy importante en la que, con el paso del tiempo, casi todos los expuestos y sólo unos pocos de los no expuestos enferman. Si se seleccionan los controles entre los participantes no enfermos al final del periodo, la proporción de expuestos ya no será del 50 %, es decir, los controles no estarán siendo un buen reflejo de la exposición de la población de origen de los casos.

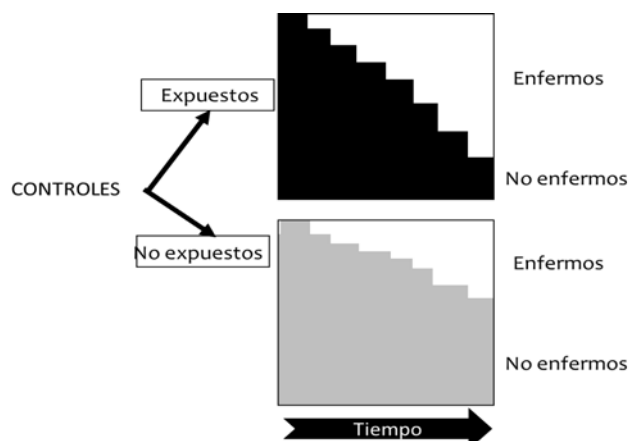
Figura 4. Muestreo acumulativo. Selección de todos los casos y una muestra de los no casos al final del periodo de riesgo.



Hay una excepción: si la enfermedad es muy rara, la diferencia en la proporción de expuestos al principio y al final del periodo será muy pequeña. En este caso, la *OR* será un buen estimador de la **razón de riesgos**.

- b) Seleccionando a una muestra de sujetos de la cohorte al inicio del seguimiento, sin tener en cuenta si esas personas, durante el seguimiento terminan siendo casos o no (**caso-cohorte**) (Figura 5).

Figura 5. Muestreo caso-cohorte. Selección de todos los casos y una muestra de los no casos de la cohorte al inicio del seguimiento.



De esta manera estaríamos aproximándonos a la definición misma del denominador de la incidencia acumulada: la población en riesgo al inicio del periodo. En este caso, **la OR permite estimar la razón de riesgos** de la cohorte original, independientemente de si la enfermedad a estudio es rara o no. Este tipo de diseños es especialmente habitual cuando, para medir la exposición, se necesita hacer determinaciones muy caras en muestras biológicas.

- c) Buscando uno o varios controles que estén a riesgo en el momento del tiempo en el que se diagnostica cada caso y asignándole la exposición que tuviera hasta ese momento (**muestreo de densidad –“incidence density sampling”, concurrente o casos y controles anidado en una cohorte**). Con este diseño, un mismo sujeto podría ser, por ejemplo, control de un caso producido en el primer año de seguimiento, y convertirse posteriormente en caso en el segundo año, según el momento del tiempo al que nos refiramos. En este diseño, si se empareja por tiempo de seguimiento, la *OR* se aproxima a la **razón de tasas** del estudio de cohortes original, de nuevo con independencia de la frecuencia del enfermar, y, además, permite controlar los posibles cambios que haya tenido la exposición en la población a lo largo del tiempo.

5.1.3. Otros diseños

Se han definido los **estudios de casos y controles de prevalencia** como aquellos diseños transversales en los que se muestrea una proporción de los controles, o de los controles y de los casos, en lugar de trabajar con todos ellos. Si los controles

se obtienen de aquellos sujetos que están sanos, la *OR* del estudio estimará la razón de odds poblacional. Si por el contrario los controles se obtienen muestreando toda la población, incluidos los que en algún momento desarrollarán la enfermedad, la *OR* será un buen estimador de la razón de prevalencias poblacional.

Los **estudios de casos cruzados**, (o case-crossover) (1) se utilizan para estudiar exposiciones en elevaciones transitorias del riesgo. Todos los sujetos de estudio son casos (personas en las que ha ocurrido la enfermedad o desenlace de interés). Se determina si se ha producido la exposición en un corto periodo de tiempo definido, inmediatamente anterior a la ocurrencia del evento (periodo de riesgo o periodo caso) y se compara con la exposición en uno o más periodos control; estos últimos tratan de obtener información sobre el nivel de exposición habitual del sujeto. De esta forma cada persona es caso y su propio control. Los desenlaces típicos que se estudian con estos diseños son enfermedades agudas de comienzo brusco y claro como infarto, ictus, caídas, accidentes de tráfico, etc. Son especialmente útiles para estudiar el efecto de posibles factores desencadenantes.

5.2. Fuente de los controles: selección primaria vs fuentes secundarias

Si los casos son representativos de una **población claramente definida** (por ejemplo, existe un registro como podría ser el padrón municipal (registro de los habitantes de un municipio) o las bases de tarjeta sanitaria, que identifican a la población) la mejor manera de seleccionar los controles es mediante muestreo aleatorio dentro de dicha población. **Cuando esto no es posible, utilizamos fuentes secundarias para la identificación de controles.** Algunos ejemplos de posibles fuentes secundarias, con aspectos específicos a tener en cuenta en cada uno, serían: (1,4)

- **Vecindario:** útil especialmente cuando los casos se obtienen de una clínica u hospital que ofrece cobertura a dicha zona, aunque tiene riesgo de emparejamiento por variables relacionadas con la residencia (i.e. exposiciones ambientales, factores socioeconómicos como la renta).
- **Amigos o conocidos:** riesgo de existir emparejamiento inadvertido para actividades que realizan juntos o para las que existe afinidad, como ocurre con exposiciones ligadas al estilo de vida. La elección del control es sugerida por el caso, con el correspondiente sesgo que pueda conllevar. Las personas más “sociables” tenderían a estar sobrerrepresentadas como controles.
- **Redes sociales:** sencillo y económico, pero es difícil entender bien a veces a qué población se está invitando a participar y hasta qué punto puede representar a la población origen de los casos. Es un entorno susceptible al sesgo del voluntario sano o de reclutar personas especialmente interesadas en un tema que quizás podría tener relación con la exposición de interés.

- **Hermanos:** se suele usar para controlar por factores genéticos, pero existe mucho riesgo de confusión por las exposiciones compartidas, incluso intraútero.
- **Controles hospitalarios:** se espera que los controles representen a la población de referencia que habría acudido al mismo hospital que los casos de haber desarrollado la enfermedad. El riesgo de que la enfermedad que se toma como referencia pueda estar inadvertidamente asociada con la exposición hace que cada vez más se tienda a seleccionar varias enfermedades como controles, todas ellas en principio no asociadas a la exposición. A la hora de elegir la lista de enfermedades que sirvan para seleccionar controles es importante tener en cuenta que deben ser independientes de la exposición de interés. Es decir, dicha exposición no debe ser factor de riesgo para estas enfermedades.
- **Controles fallecidos:** adecuados en diseños de mortalidad proporcional, siempre y cuando los controles no fallezcan de enfermedades relacionadas con la exposición a estudio.

En algunos estudios se seleccionan **más de un tipo de controles por cada caso**. Es útil si con cada grupo pretendemos probar una **hipótesis diferente, aunque complementaria**, por ejemplo: en un diseño en el que reclutamos un grupo de controles sanos y un grupo de controles que estén enfermos de otra patología diferente a la que se va a estudiar, para comprobar que los resultados no se deben exclusivamente a un sesgo de memoria diferencial entre sujetos sanos y enfermos.

Otra posibilidad es comparar la exposición en los casos con aquella teórica esperada en la población general, especialmente útil en los estudios genéticos.

5.3. Ratio de controles por caso

Puesto que los estudios de casos y controles suponen un muestreo dentro de la cohorte original, suelen suponer una pérdida de precisión en las estimaciones con respecto a las de la cohorte de referencia. Si reclutar casos es más caro, complejo o simplemente, no los hay o no están disponibles, una forma eficiente de poder alcanzar **el tamaño muestral requerido** es reclutando **varios controles por cada caso**. En general, pasar de 4 controles por caso ya no supone un aumento importante de la potencia, aunque puede ser una estrategia útil para exposiciones muy raras, *OR* grandes, estudios que manejan valores alfa muy pequeños (por ejemplo, estudios de asociación genética), o cuando el coste de incorporar más controles en los análisis es muy pequeño (información ya disponible) (6).

6. La medida de la exposición

La definición de la exposición y su medida son aspectos fundamentales en cualquier estudio epidemiológico. El término “exposición” se suele referir a una característica externa o ambiental que afecta al individuo, pero se usa de forma genérica para referirse a cualquier tipo de determinante —variable independiente, hábito, característica personal— que es el objeto principal de estudio. Puede ser, por ejemplo, la ciudad de nacimiento, el lugar del cuerpo en el que tiene lunares, el uso de pesticidas, el gusto por lo picante o tener una determinada variante genética.

6.1. La definición de la exposición

En los estudios de casos y controles, muchas veces, definir la exposición entraña dificultades específicas porque estamos buscando información sobre eventos o cosas que ocurrieron en el pasado. El primer paso en esta tarea es definir teórica y, sobre todo operativamente, de acuerdo con nuestra hipótesis causal, cuál es nuestra exposición de interés con tres conceptos básicos: qué, cuándo y cuánto. Si seguimos con el ejemplo del café, tendremos que definir aspectos como el tipo de café, qué entendemos por desayuno, la cantidad consumida en el desayuno o cuántas veces a la semana se consume; por ejemplo: podríamos definir como expuestos a los sujetos que consuman al menos una taza de café no descafeinado, mínimo dos veces por semana en la primera comida del día. En este marco, es esencial tener en cuenta el tiempo de exposición y cómo definimos la presencia e intensidad de la dosis de exposición. (1)

- **Tiempo.** Cuando hablamos de tiempo, en este contexto nos referimos a dos conceptos diferentes:
 - ♦ Por una parte, es relevante el ***momento en el que se produce la exposición***. Por ejemplo, si quiero conocer si el uso del teléfono móvil durante la conducción es un factor de riesgo para tener accidentes de tráfico, me interesa saber si la persona ha estado usando el móvil mientras conducía en el momento que tuvo el accidente, no nos vale que fuera una semana antes. También hay exposiciones cuyo efecto biológico depende de la etapa de la vida en la que se produce (periodos vulnerables).
 - ♦ Por otro lado, son importantes la ***frecuencia y duración de las exposiciones***. Hay relaciones causales que implican exposiciones crónicas y/o en las que se sabe que hay un periodo de tiempo largo desde que se produce la exposición hasta que la enfermedad se inicia o se diagnostica (periodos de inducción o latencia). Es, por ejemplo, el caso de la relación

entre tabaco y cáncer de pulmón, en el que es más relevante lo fumado en los años anteriores que el estatus de fumador en el momento del reclutamiento. De hecho, como se desconoce el momento de inicio de la enfermedad (que es anterior al momento en que es diagnosticable), se suele descartar la exposición de un periodo variable próximo al diagnóstico (por ejemplo, la exposición del último año anterior al diagnóstico e ídem en los controles, pero teniendo como referencia el momento de selección como control).

- **Dosis.** Como en el caso anterior, es imprescindible conocer bien el campo de estudio en el que se trabaja para poder hacer una definición operativa de la dosis de interés útil para la hipótesis que se plantea, bien sea a través de variables dicotómicas (sí/no), ordinales (bajo, medio, alto) o cuantitativas (dosis media, dosis máxima, dosis acumulada, etc.).

Muchas veces, lo más adecuado es combinar la información de la dosis y del tiempo de exposición. Esto es especialmente relevante en aquellos casos en que pensamos que la exposición se acumula en el organismo (dosis acumulada). Un ejemplo que se utiliza con mucha frecuencia en la literatura sería expresar la exposición al tabaco en paquetes-año ($n.^{\circ}$ de cigarrillos diarios $\times$ $n.^{\circ}$ de años/20). Otro ejemplo podría ser un estudio con casos de cáncer de piel y controles sanos en el que quiero estudiar, como posible factor de riesgo, la exposición al sol. Podría considerar como expuestas a las personas que han pasado más de un determinado número de horas semanales expuestas al sol, aunque también podría interesarme la dosis acumulada a lo largo de la vida o la dosis máxima recibida durante la infancia, como periodo de especial vulnerabilidad.

6.2. Forma de medida de la exposición

En los estudios de casos y controles clásicos, se busca información de exposiciones que han ocurrido en el pasado. Es bastante habitual que la medición de la exposición se realice preguntando al propio sujeto sobre ella, por ejemplo, mediante el uso de *cuestionarios*, lo que, lógicamente, implica que puede estar sujeta a sesgos de recuerdo. No es, sin embargo, la única alternativa. A veces se dispone de información de exposición a partir de *variables registradas* cuando esta ocurrió, por ejemplo, en la historia clínica del paciente. Otra posibilidad es utilizar información ligada al entorno del sujeto (i.e. exposición a contaminantes según el lugar de residencia del participante, sea caso o control). También es cada vez más frecuente el uso de *biomarcadores* que se relacionan con la exposición de interés. En este caso, cuando se trata de enfermedades crónicas o con largos periodo de latencia, es importante considerar hasta qué punto la propia enfermedad o el tratamiento aplicado para ella han

podido modificar los valores del biomarcador utilizado, es decir, tenemos riesgo de encontrar asociaciones debidas a causalidad inversa. Este problema no existe en los diseños de casos y controles anidados o caso-cohorte, ya que las muestras biológicas se recogen para todos antes de que ocurra la enfermedad.

Por último, es relativamente frecuente que, en aquellos casos en los que la enfermedad es rara y la exposición difícil de medir, se asuma que la exposición de los controles es equivalente a la exposición media poblacional. Por ejemplo, en un estudio sobre cáncer y exposición a metales pesados podríamos medir la concentración de metales solo en los casos, si se trata de una técnica muy costosa, y asumir que los controles tendrán valores similares a la población general.

7. Medidas de asociación

Como ya se ha visto en el capítulo correspondiente, en los estudios de casos y controles la principal medida de asociación que se utilizan son las odds ratio. En nuestro caso, comparamos la odds de exposición, que es la razón entre la probabilidad de estar expuesto y la probabilidad de no estarlo. Un estudio de casos y controles nos permite calcular hasta qué punto difieren la odds de exposición en casos y en controles. Teóricamente, ese no es el parámetro que nos interesa: en todo caso, querríamos comparar la odds de enfermar en expuestos y la odds de enfermar en no expuestos. Sin embargo, en la práctica la razón de odds de exposición y la de enfermar son equivalentes.

Población	Enfermos	Sanos
Expuestos	A	B
No expuestos	C	D

En un estudio de cohortes podríamos calcular la odds de enfermar en expuestos $= \frac{A}{B}$ y la odds de enfermar en no expuestos $= \frac{C}{D}$; la razón de odds sería $OR = \frac{AD}{BC}$. En un estudio de casos y controles, lo que podemos calcular es la odds de exposición en casos $= \frac{A}{C}$ y la odds de exposición en controles $= \frac{B}{D}$; la razón de odds sería, de nuevo, $OR = \frac{AD}{BC}$. Es decir, la razón de odds es la misma, tanto si la calculamos partiendo de la probabilidad de enfermar como de la probabilidad de

estar expuesta. Los estudios de casos y controles, por lo tanto, van a permitirnos siempre aproximarnos a la razón de odds poblacional. La odds ratio, de cualquier forma, es una medida de interpretación complicada, pero tiene la ventaja de que, en determinadas circunstancias, es también un estimador insesgado de otras medidas de asociación:

- En estudios con una adecuada selección de controles por muestreo acumulativo la *OR* sería un estimador insesgado de la razón de riesgos de esa población y estimarían la razón de riesgos poblacional si la enfermedad es infrecuente. (7)
- En casos-cohorta la *OR* es un estimador insesgado del riesgo relativo poblacional, como ya hemos comentado.
- En estudios con muestreo de controles por densidad, la *OR* es un estimador insesgado de la razón de tasas poblacional.

8. Confusión y sesgos en estudios de casos y controles

La identificación de posibles variables confusoras y sesgos debe realizarse de manera individualizada para cada estudio, puesto que dependen de su diseño y las circunstancias en las que se desarrolla. Existen técnicas que permiten tener en cuenta y corregir el efecto de las variables confusoras en las odds ratios como son el análisis estratificado y el uso de modelos multivariantes que, en el caso de los estudios de casos y controles, son principalmente los modelos de regresión logística (ver capítulos 5 y 6).

Aunque no hay reglas generales que se ajusten a todas las situaciones, sí podemos recordar algunos sesgos que típicamente se relacionan con los diseños de casos y controles: (1,8)

Sesgos en la selección de casos: Por la selección de casos prevalentes (sesgo de supervivencia): seleccionar los casos entre sujetos de registros de casos, consultas, hospitalización, etc. tiene la ventaja de que no hay que esperar a que aparezcan nuevos casos de enfermedad (casos incidentes) para reclutarlos. Sin embargo, esto puede hacer que perdamos la posibilidad de incluir en el estudio a los casos con menor supervivencia y que algunas de las variables que caractericen a los casos tenga más que ver con la probabilidad de reclutarlos (i.e. es más fácil encontrar a aquellos que requieren ingresos hospitalarios frecuentes) que con el propio enfermar.

Sesgos en la selección de controles: Selección a partir de una población o subpoblación que no es la misma o no representa bien a la población que dio origen a los casos. A veces se incurre en este error por tratar que los controles sean muy similares a los casos. Por ejemplo, pedir a los casos que recluten entre familiares o amigos a los controles para que compartan variables de tipo sociodemográfico. Puede ser que también compartan la distribución de algunas exposiciones, incluidas las de las variables de estudio, de manera diferente a como se distribuyen en la población general.

Sesgos de clasificación: En aquellos casos en los que la información sobre la variable de exposición se recoja de manera retrospectiva, es posible que el hecho de tener o no tener la enfermedad influya en la exactitud con la que se recuerda la exposición que estamos estudiando. Habitualmente nos referimos a esto como sesgo de recuerdo.

9. Emparejamiento

Una de las herramientas con las que se cuenta en los estudios de casos y controles para mejorar la eficiencia es el emparejamiento entre casos y controles. Sin embargo, a pesar de lo que muchas personas creen, el emparejamiento no elimina la confusión relacionada con las variables por las que se empareja, es decir, en un estudio caso-control emparejado siempre hay que tener en cuenta en el análisis los factores de emparejamiento.

El emparejamiento puede permitirnos tener en cuenta variables confusoras, tanto observadas y medidas, como aquellas que no recogemos en el estudio, pero se asocian con estas. También tiene la ventaja de que facilita el análisis dentro de determinados subgrupos, por ejemplo, asegurándonos que hay suficiente representación entre los controles de sujetos de una determinada edad para que se puedan calcular con facilidad las razones específicas. Muchas exposiciones se asocian con la cohorte o generación de pertenencia —las mujeres muy jóvenes—, por ejemplo, han podido estar expuestas a los anticonceptivos, fecundación in vitro, etc., una oportunidad que no estaba disponible para mujeres muy mayores. Si los controles seleccionados aleatoriamente resultan mucho más jóvenes que los casos, en las celdillas extremas de edad podría haber únicamente controles —muy jóvenes— o casos —casos de edad más avanzada—, con la consiguiente pérdida de potencia estadística. Por ello, emparejar por edad y sexo es una práctica muy común en los

estudios de casos y controles. No obstante, el emparejamiento puede tener sus riesgos y ser origen de sesgos si no se reflexiona cuidadosamente sobre él en la fase de diseño, y, durante la fase de análisis, es importante ajustar siempre por las variables de emparejamiento, para evitar que actúen como confusoras, independientemente del motivo por el cual se apareó. (9)

Las estrategias de emparejamiento más habituales son: (1)

- De manera **individual** (datos emparejados) en base a variables confusoras predefinidas, es decir, haciendo que cada control sea similar a uno de los casos en cuanto a determinadas características como un rango de edad, sexo, etc. Esta circunstancia se da, por ejemplo, cuando se seleccionan controles en el momento del diagnóstico del caso correspondiente (emparejamiento por densidad de incidencia). En estas situaciones, en las que un control o grupo de controles sólo pueden servir para un caso determinado, se requiere mantener la relación de emparejamiento en el análisis.
- De manera **grupal**, encontrando un **grupo (emparejamiento de frecuencia)** de controles que tengan distribuciones similares a los casos en cuanto a las variables a estudio, por ejemplo, buscando un grupo de controles que tengan la misma proporción de hombres y mujeres que los casos. En este caso, lo que es imprescindible es tener en cuenta la variable de emparejamiento, normalmente introduciendo dicha variable o variables como variables de ajuste en el modelo multivariante.

Dentro de este punto, es importante señalar el riesgo de sobreemparejamiento: si se empareja por una variable que no es factor de confusión, estamos obligados a controlar esa variable en el análisis y no ganamos en eficiencia. En determinadas circunstancias, sobreemparejar supone perder la capacidad de estudiar la exposición de interés. Esta situación se produce cuando emparejamos por una variable muy correlacionada con la exposición o una variable que representa un paso intermedio entre la exposición y la enfermedad (i.e. buscar controles entre los vecinos de los casos hace que no puedas estudiar exposiciones ambientales ligadas al lugar de residencia).

10. Ejemplos de estudios de casos y controles

- Doll R, Hill AB. *Smoking and Carcinoma of the Lung*. Br Med J. 1950; 2(4682): 739-748. <https://doi.org/10.1136/bmj.2.4682.739>.

Al comienzo del siglo XX se detecta un aumento en los casos de cáncer de pulmón. La opinión pública está convencida de que se relaciona con la automovilización e industrialización de la sociedad. Este estudio, convertido ya en un clásico, demuestra la importante asociación entre el consumo de tabaco y el cáncer de pulmón.

- Herbst A L, Ulfelder H, Poskanzer DC. *Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women*. N Engl J Med. 1971;284(15):878-81. <https://doi.org/10.1056/NEJM197104222841604>.

Otro estudio de casos y controles clásico. Se había observado una agrupación (o clúster) de 8 pacientes jóvenes que tenían un adenocarcinoma vaginal de células claras, un tumor muy infrecuente. Por cada paciente se seleccionaron cuatro controles, emparejadas a los casos por día, hospital de nacimiento y tipo de servicio (sala o privado), y se encuestó a las madres sobre lo sucedido durante la gestación. Este estudio demostró la asociación entre el tratamiento con dietilestilbestrol durante el embarazo —un medicamento que se daba para reducir riesgo de abortos—, y la aparición de dicho cáncer.

- Castaño-Vinyals G, Aragonés N, Pérez-Gómez B *et al.* & MCC-Spain Study Group. *Estudio multicaso-control de base poblacional de tumores comunes en España (MCC-Spain): razón y diseño del estudio*. Gac Sanit. 2018;32(5):501.

<https://doi.org/10.1016/j.gaceta.2018.04.004>.

En este artículo se presenta el diseño de MCC-Spain, un ambicioso estudio multicaso-control español para estudiar la influencia de factores genéticos y ambientales en el origen de varios de los principales tipos de cáncer en España (mama, colon, próstata, estómago, leucemia linfática crónica). Usa un mismo grupo de controles para todos los tipos de casos (teniendo en cuenta, obviamente, el sexo). El protocolo del estudio está también publicado en:

https://mccspain.org/wp-content/uploads/2016/04/01_PROTOCOLO_ESTUDIO_MCC_v16.pdf

- Castelló A, Pollán M, Buijsse B *et al.* *Spanish Mediterranean diet and other dietary patterns and breast cancer risk: case-control EpiGEICAM study*. Br J Cancer. 2014;111(7):1454-62. <https://doi.org/10.1038/bjc.2014.434>.

Estudio de casos y controles emparejados de manera individual sobre patrones dietéticos y cáncer de mama utilizando como controles mujeres

amigas, conocidas o familiares no relacionadas genéticamente con los casos. Por tanto, el análisis se realiza teniendo en cuenta el emparejamiento.

- García-Fulgueiras A, Navarro C, *et al.* *Legionnaires' disease outbreak in Murcia, Spain*. Emerg Infect Dis 2003;9(8):915-21. <https://doi.org/10.3201/eid0908.030337>.

Por último, un ejemplo de un uso muy clásico de los estudios de casos y controles: el análisis y búsqueda del origen de un brote, en este caso uno de los mayores brotes de legionelosis que ha habido en el mundo y que se registró en Murcia.

11. Referencias

1. Celentano D.D, Szklo M. *Gordis. Epidemiología*, 6ª edición. Ed. Elsevier. Barcelona, 2019.
2. Vandenbroucke JP, Pearce N. *Case-control studies: basic concepts*. Int J Epidemiol. 2012 Oct;41(5):1480-9.
3. Vandenbroucke JP, Pearce N. *Incidence rates in dynamic populations*. Int J Epidemiol 2012;41:1472-1479.
4. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4th edition. Ed. Wolters- Kluwer/ Lippincott Williams and Wilkins. Mexico, 2021.
5. Westreich D. *Epidemiology by design. A causal approach to the health sciences*. Ed. Oxford University Press. New York, 2020.
6. Katki HA, Berndt SI, Machiela MJ, *et al.* *Increase in power by obtaining 10 or more controls per case when type-1 error is small in large-scale association studies*. BMC Med Res Methodol. 2023 Jun 29;23(1):153.
7. Pastor-Barriuso R. *Bioestadística*. 2012.
<http://hdl.handle.net/20.500.12105/5420>
8. Schlesselman, JJ. *Case- Control Studies. Design, Conduct, Analysis*. Oxford University Press. 1982.
9. Pearce N. *Analysis of matched case-control studies*. BMJ 2016;352:969.

Capítulo 12.

Ensayos por conglomerados

ISABEL DEL CURA GONZÁLEZ y JESÚS MARTÍN FERNÁNDEZ

- 1. Introducción**
- 2. Unidad de aleatorización, unidad de intervención y unidad de análisis**
- 3. Aleatorización y cegamiento**
- 4. Muestra y muestreo: el efecto diseño**
- 7. Análisis estadístico**
- 6. Intervenciones complejas**
- 7. Aspectos éticos en los ensayos por conglomerados**
- 8. Como reportar con calidad un ensayo por conglomerados**
- 9. Referencias**

1. Introducción

Los ensayos por conglomerados son una modalidad de ensayo con asignación aleatoria en los que la unidad de aleatorización son grupos en vez de individuos. El término conglomerado hace referencia al agrupamiento, y la unidad a aleatorizar es el grupo (comunidades, hospitales, escuelas, instituciones, etc.). Se caracterizan en la mayoría de los casos por su naturaleza multinivel, siendo la situación más habitual la inclusión de dos niveles: el nivel del conglomerado y el nivel de sus miembros individuales, aunque pueden existir ensayos de más de dos niveles (1).

Son comunes en la investigación en servicios de salud, en salud pública, en promoción de la salud y en epidemiología clínica cuando evaluamos la efectividad o la implementación posterior de intervenciones complejas. Son útiles para ayudar a comprender qué efecto podría tener la introducción de una nueva tecnología o estrategia sanitaria en la salud de la población, informar la toma de decisiones e investigar sobre intervenciones complejas que incluyen las no farmacológicas como cambios en el estilo de vida, tipos de atención o intervenciones educativas. (2,3,4) En los últimos años estamos asistiendo a un interés creciente por este tipo de ensayos, incorporándose incluso en estudios basados en datos de historia clínica (5) y en la investigación en salud global, con una creciente popularidad también en los países de ingresos bajos y medios. (6)

Estos diseños son especialmente útiles cuando los sujetos no pueden ser asignados de forma individual o cuando pueden interactuar entre sí durante el período que reciben la intervención. La cuestión clave es saber cuándo son la mejor opción, (7) teniendo en cuenta que son menos eficientes desde el punto de vista estadístico y suelen ser más complejos desde el punto de vista logístico que los ensayos con aleatorización de individuos, por lo que es fundamental revisar la justificación de su uso. En líneas generales, se aleatorizan grupos de individuos en lugar de individuos para minimizar el riesgo de contaminación, estudiar intervenciones que se aplican a grupos, utilizar eficazmente recursos limitados o resolver problemas logísticos y administrativos. (8) Por ejemplo, si evaluamos la efectividad de intervenciones dirigidas a capacitar a los profesionales o cambios de procesos o de tecnologías utilizadas en determinados centros sanitarios, es muy difícil realizar una asignación aleatoria a nivel individual (p. ej. profesionales de la salud), ya que en algunos casos todos los profesionales de un mismo centro o servicio estarán expuestos a la intervención en cuestión. En otros casos, si evaluamos un programa de educación sanitaria en centros escolares, los alumnos pueden influirse mutuamente, y no pueden considerarse independientes unos de otros, por lo que, para

evitar ese problema, las escuelas pueden ser asignadas aleatoriamente a recibir el programa educativo o actuar como control. Cuando hay riesgo de contaminación, el diseño de asignación aleatoria por conglomerados puede ser mejor que el de asignación individual. En la tabla 1 se detallan algunos de los escenarios más habituales en los que se realizan ensayos por conglomerados.

Tabla 1. Escenarios adecuados para un diseño de aleatorización por conglomerados

Escenario	Ejemplo
Cuando la intervención debe aplicarse al grupo.	Efecto de la fluoración del agua de la red de suministro público sobre la salud bucodental (no se puede distribuir aleatoriamente a ciudadanos individuales a que les llegue el agua fluorada o no, sino que deben asignarse aleatoriamente pueblos, ciudades...).
Cuando la intervención implica entrenamiento o educación de profesionales sanitarios para evaluar la repercusión en el conjunto de sus pacientes.	Entrenamiento a una serie de psicólogos en una determinada técnica de psicoterapia grupal para compararla con otra. Se asignan aleatoriamente diferentes centros de salud mental, de forma que unos realizarán la terapia de grupo de un modo y el resto de otro, para valorar su repercusión en el conjunto de los pacientes.
Cuando se desea disminuir el efecto de la contaminación de la intervención.	Campaña informativo-educativa en centros de salud sobre el lavado de manos en la prevención del contagio del resfriado común. Es previsible que los pacientes hablen entre ellos en la sala de espera, al salir de la consulta, etc.; por lo que, si dentro de ese centro de salud unos recibieran la intervención y otros no, los resultados estarían contaminados ya que aun sin recibir la intervención podrían conocerla e incluso practicarla, lo que afectaría a los resultados.
Cuando se desea evaluar el efecto de la intervención tanto a nivel individual como grupal.	Efecto de un determinado programa de vacunación. Se asignan aleatoriamente diferentes áreas sanitarias o diferentes ciudades, provincias, etc., a recibir la vacuna o no, y se evaluarían tanto el efecto y la seguridad sobre el individuo como la efectividad (repercusión real sobre el conglomerado: inmunidad colectiva, incidencia de infección en la comunidad, etc.).

Tomada de Prieto E, Portoles A. (8)

Los ensayos por conglomerados en general son estudios de corte pragmático en los que se evalúa la efectividad más que la eficacia. Se estudia el potencial efecto beneficioso de una intervención en las condiciones habituales de la práctica asistencial, preventiva o de promoción de la salud, buscando ser altamente representativos de las poblaciones en su contexto habitual. Lo que interesa es conocer cómo funciona la estrategia en su conjunto (fármacos, pruebas diagnósticas, consejo clínico preventivo u otro tipo de intervenciones) y no aislar la “eficacia” teórica, en condiciones ideales, de cada una de sus partes. Esto hace que tengan una gran aplicabilidad en la denominada investigación en efectividad comparada —*comparative effectiveness research*

CER—, (9) constituyendo uno de los diseños útiles para generar información para la toma de decisiones cuando se investiga sobre intervenciones complejas o cuando el cegamiento no es posible. Por otro lado, en este tipo de diseños se suelen incorporar medidas de resultado que incorporan la perspectiva del paciente o PROMS (*Patient-reported outcomes*). (10) Herramientas como PRECIS (*Pragmatic Explanatory Continuum Indicator Summary*), que nos ayudan a identificar los elementos diferenciales de un enfoque explicativo y un enfoque pragmático en los ensayos (11).

A continuación, vamos a revisar algunos de los aspectos metodológicos diferenciales de los ensayos por conglomerados. Aclarar previamente dos conceptos que pueden inducir a error por la terminología utilizada:

- a) Los ensayos aleatorios por conglomerados no tienen relación con el análisis de clúster, una técnica estadística exploratoria multivariante utilizada para definir grupos de personas similares, y
- b) las cuestiones estadísticas que plantean los ensayos aleatorios por conglomerados son diferentes de las que plantea el muestreo por conglomerados, en el que se identifican conglomerados naturales como las zonas geográficas, y se eligen algunos conglomerados para ser estudiados, preferentemente al azar.

2. Unidad de aleatorización, unidad de intervención y unidad de análisis

En los ensayos por conglomerados, la unidad de aleatorización, la unidad de intervención y la unidad de análisis pueden ser diferentes. La unidad de aleatorización es un grupo social cuyos componentes comparten una característica común (vivir en un pueblo, pertenecer a un área sanitaria, ser asistidos en un centro sanitario, acudir a un instituto, ser atendidos por un mismo profesional, etc.); la unidad de intervención podrá ser el grupo en su conjunto (una campaña de salud nacional en medios de comunicación, en la que la unidad de intervención será el conjunto de ciudadanos de ese país) o los individuos que forman el grupo (un programa de vacunación, en el que la unidad de intervención es cada uno de los individuos que recibe la vacuna); y la unidad de análisis puede ser el conglomerado o cada uno de sus individuos. A continuación, revisaremos algunos ensayos por conglomerados para identificar las diferentes posibilidades entre estas unidades.

En el ensayo MULTIPAP, (12,13) cuyo objetivo fue evaluar la efectividad de una intervención en profesionales que mejore la adecuación de la prescripción en pacientes con multimorbilidad y polimedicación, los médicos de familia fueron asignados al grupo intervención, que recibió la intervención eMULTIPAP (un programa intensivo de educación), (14) o al grupo control, que seguía la práctica habitual. Se seleccionó una muestra de los pacientes de cada médico de familia y todos los pacientes incluidos por un mismo médico formaban parte de un mismo clúster o conglomerado. En este ejemplo, la unidad de aleatorización y la unidad de intervención fue el médico (proveedor de la atención), y la unidad de análisis el paciente en el que se medía el resultado (adecuación de la prescripción y calidad de vida). En este caso el conglomerado es el médico y no el centro porque la prescripción es un factor muy dependiente del profesional. En el estudio CUIDACARE, (15) que evalúa la efectividad de una intervención domiciliaria realizada por enfermeras sobre la calidad de vida de los cuidadores de personas con discapacidades o enfermedades crónicas que viven en la comunidad, la unidad de aleatorización fueron los centros de salud, la unidad de intervención las enfermeras (todas las enfermeras de un mismo centro recibieron la misma intervención dirigida a cuidar al cuidador) y la unidad de análisis las personas cuidadoras de pacientes inmovilizados en domicilio. En el estudio PROLACT, (16) que evalúa la efectividad de una intervención educativa grupal implementada por enfermeras en los centros de salud para promover la lactancia materna exclusiva hasta los 6 meses, la unidad de aleatorización fueron los centros de salud, y la unidad de intervención y análisis fueron las díadas mujeres y niños. Distribución similar tiene el estudio EDUCORE (17) para informar sobre el riesgo cardiovascular en pacientes con hipertensión arterial no controlada, o el que compara intervenciones para ayudar a dejar de fumar. (18)

En los ensayos por conglomerados se deben considerar dos conjuntos de criterios de elegibilidad; siguiendo con el ejemplo del estudio MULTIPAP, hay unos criterios de elegibilidad de los grupos (médicos de familia) que se incluirán en el ensayo y otros criterios de elegibilidad para los participantes individuales que se incluirán en los grupos o clústeres (por ejemplo, los pacientes elegibles dentro de la consulta de cada médico de familia debían tener multimorbilidad).

3. Aleatorización y cegamiento

Clásicamente en los ensayos por conglomerados se aleatorizan estos de forma convencional en grupos paralelos, (1) y se puede utilizar para ello un método aleatorio simple, un diseño de conglomerados emparejados o un diseño estratificado.

Si se utiliza un factor de emparejamiento adecuado, el diseño emparejado gana potencia en relación con el diseño completamente aleatorio. Sin embargo, estas variables de emparejamiento pueden ser difíciles de identificar, especialmente si el número de conglomerados a emparejar es grande. El cálculo del coeficiente de correlación intraclúster (CCI) a partir de un diseño emparejado suele ser problemático porque la variación entre clústeres no puede separarse de los efectos de la intervención o las intervenciones y de los factores de emparejamiento, por lo que suele ser preferible el uso de diseños estratificados.

En un ensayo aleatorio por grupos, la asignación de la intervención está predeterminada para cada miembro del clúster, se asignan grupos en lugar de individuos, y el número de grupos suele ser pequeño, por lo que el riesgo de que puedan producirse desequilibrios por azar en los factores pronósticos es mayor que en los ensayos con aleatorización de individuos, y también se puede producir cuando el número de conglomerados es grande. Este desequilibrio en determinadas características individuales también puede deberse a sesgos de selección. (19)

En los ensayos con aleatorización de individuos se recomienda que los participantes y los evaluadores de resultados estén cegados a la intervención recibida siempre que sea posible. Sin embargo, en los ensayos aleatorizados por grupos, el cegamiento es más complicado. A menudo, la ejecución de las intervenciones en los ensayos por conglomerados implica a una serie de personas diferentes (por ejemplo, formadores de los profesionales de la salud en el uso de las directrices, profesionales de la salud que aplican la intervención a sus pacientes, pacientes que reciben intervenciones en condiciones de práctica clínica, evaluadores de resultados que pueden ser de profesionales de salud, pacientes o un grupo completamente diferente). El cegamiento de la intervención a menudo es imposible, aunque en algunos casos es posible el cegamiento de los que evalúan el resultado. Es imprescindible indicar explícitamente en el informe del ensayo, si hubo o no cegamiento, quién estaba o no cegado y, si no se realizó cegamiento, los motivos. (1)

El ensayo por grupos de cuña escalonada (*stepped wedge cluster randomised trial*, SW-CRT) es un tipo de ensayo por conglomerados que se utiliza cada vez más, especialmente para evaluar intervenciones de prestación de servicios sanitarios u otras intervenciones a nivel de conglomerados. (20) Este diseño se caracteriza por el hecho de que los conglomerados se asignan al azar a varias secuencias diferentes que dictan el momento en que el conglomerado cambiará de la condición de control a la condición de intervención. Su atractivo fundamental es que todos los conglomerados reciben finalmente la intervención, lo que puede aumentar la aceptabilidad de una evaluación aleatoria y el atractivo social del estudio, y permite una evaluación en el contexto de una implementación rutinaria. (21) Este diseño se suele utilizar para evaluar cómo funcionarían las intervenciones en el mundo

real, con criterios de exclusión limitados, y es más pragmático que explicativo. Si la razón principal para desplegar la intervención es que ejerza sus beneficios, y si la implantación por fases es inevitable, un ensayo de cuña escalonada es una opción válida y proporciona mejores evidencias que la mayoría de las evaluaciones no aleatorias. Sin embargo, el SW-CRT presenta un mayor riesgo de sesgo y requiere que los investigadores proporcionen una justificación clara de la elección del diseño del estudio. (22-25) Son una opción para llevar a cabo una evaluación aleatoria que de otro modo no sería posible, (26) pero la mera popularidad y novedad de la SW-CRT no debería ser un factor para su adopción. En las situaciones en las que un ensayo aleatorio por conglomerados paralelo convencional es factible, es probable que sea el diseño preferido.

4. Muestra y muestreo: el efecto diseño

En los ensayos por conglomerados las observaciones tienden a ser más parecidas entre los participantes de un mismo conglomerado que entre los de dos conglomerados diferentes. Estos diseños presentan dos elementos de variabilidad, una interconglomerado y otra intraconglomerado, que deben tenerse en cuenta en el análisis estadístico y en el cálculo del tamaño muestral. (27) Debido a la variabilidad intergrupos y a la “no independencia” de las observaciones resultantes, la aleatorización por conglomerados posee menor poder estadístico y es menos eficiente que la aleatorización de individuos, incrementándose el error en la estimación del efecto de la intervención, ampliándose los intervalos de confianza y aumentando los valores *p* en comparación con un estudio del mismo tamaño que utilice aleatorización de individuos. (28) Para obtener la misma potencia estadística, el tamaño de la muestra debe ser mayor en un ensayo aleatorio por conglomerados que en un ensayo aleatorio a nivel individual.

El enfoque más directo para determinar el tamaño de la muestra es realizar primero el cálculo como si el diseño fuera aleatorio a nivel individual y luego inflar el valor resultante multiplicando por un factor denominado “efecto del diseño”, para proporcionar un tamaño de muestra de igual potencia que en la aleatorización por individuos. El efecto diseño cuantifica el grado en que las respuestas dentro de un conglomerado son más similares entre sí que las respuestas en conglomerados diferentes. Para calcularlo es necesario estimar primero el CCI, que informa de la proporción de la variabilidad total de los datos que puede atribuirse a las diferencias

entre conglomerados. El CCI tradicional representa la correlación entre respuestas de individuos del mismo conglomerado en un momento determinado. Otro posible CCI sería el que representa la correlación entre respuestas de un grupo de individuos en diferentes momentos en el tiempo. (29)

Si m es el tamaño del conglomerado (asumiendo que es el mismo para todos los conglomerados), entonces el factor de inflación o “efecto de diseño” asociado a la aleatorización por conglomerados es $1+(m-1) * \rho$, siendo ρ el coeficiente de correlación intraclase. Aunque normalmente ρ es pequeño (a menudo $< 0,05$), suele conocerse cuando se diseña el ensayo, ya que habitualmente se obtiene de una referencia de la literatura, y solo se estima con error después de que se haya completado el ensayo, su impacto en el tamaño de la muestra es muy importante. Cuanto mayor es el CCI, mayor es el efecto del diseño y, por tanto, mayor el tamaño muestral necesario para igualar la potencia del ensayo por conglomerados a la de un ensayo por individuos.

Para ilustrar este punto describimos el cálculo del tamaño muestral en el estudio EDUCORE:

“El tamaño de la muestra se determinó contemplando un aumento del 15 % en el número de pacientes que muestran un buen control de la hipertensión arterial en el grupo intervención frente al grupo que recibe la atención habitual, para un valor α de 0,05, y para una potencia del 80 %. Según los datos de los estudios españoles, el buen control de las presiones arteriales sistólica y diastólica se sitúa en torno al 41,4 % de los pacientes. El tamaño mínimo de la muestra se corrigió por el efecto del diseño asumiendo un coeficiente de correlación intraclase de 0,03 (30) y un tamaño medio del grupo de 30 pacientes. El tamaño final de la muestra requerida fue de 736 pacientes, asumiendo un 10 % de pérdidas en el seguimiento durante el periodo de estudio”.

Como se puede ver en el ejemplo, el tamaño de la muestra depende del tamaño medio de los conglomerados y del grado de correlación dentro de los grupos. Aunque los ensayos con un gran número de conglomerados pequeños son más eficientes desde el punto de vista estadístico que los que tienen unos pocos conglomerados grandes, estos últimos pueden ser más factibles. Además, si se van a comparar los resultados entre centros individuales, se necesitará un tamaño de muestra suficiente para lograr una precisión adecuada dentro de cada centro. En general, la potencia se incrementa más fácilmente aumentando el número de conglomerados que el tamaño de estos. Se recomienda que el número mínimo de conglomerados por brazo para asegurar un análisis válido sea al menos de cuatro, (1) y que en los cálculos del tamaño de la muestra se valore un análisis de sensibilidad, ya que las aportaciones recogidas de la literatura pueden carecer de precisión.

En los SW-CRT el cálculo del tamaño muestral es más complejo y tiene algunas peculiaridades que no son objeto de este capítulo; una buena aproximación a esta cuestión se puede revisar en el trabajo de Wang J *et al.* (31)

7. Análisis estadístico

El hecho de que el efecto de una intervención se evalúe a nivel de grupo o de participantes individuales tiene implicaciones para el análisis. Un aspecto clave de los ensayos por conglomerados, como hemos comentado, es que no se puede asumir que las observaciones realizadas sobre cada uno de los individuos que forman parte del conglomerado son independientes (a diferencia de lo que ocurre en los ensayos clásicos), por lo que es más probable que los individuos de un mismo conglomerado muestren resultados parecidos. Una preocupación importante en el análisis primario de este diseño es el uso de métodos estadísticos apropiados para dar cuenta de la correlación entre los resultados de un grupo/clúster particular. (32,33)

Los métodos clásicos de análisis, por ejemplo, los que utilizan como método de estimación los mínimos cuadrados, requieren, para su correcta especificación, que se cumplan varios supuestos, entre ellos el de homogeneidad de la varianza (homocedasticidad). Cuando la varianza de la distribución de una variable de interés no es homogénea entre los grupos, como ocurre en el caso que nos ocupa, estamos ante la presencia de heterocedasticidad. Cuando aplicamos estos métodos de análisis clásicos a los ensayos clínicos por clústeres, a pesar del incumplimiento de la asunción de homocedasticidad, pueden ofrecer estimadores del efecto de la intervención consistentes e insesgados. Sin embargo, los estimadores de la varianza y los errores estándar no lo son. El análisis que asume la homogeneidad de la varianza la infraestima en el caso de los ensayos por clústeres, ofreciendo resultados estadísticos que aumentan el error tipo I. (34)

Para evitar estos problemas se han intentado diversos abordajes. El más elemental consiste en la estimación de errores estándar robustos. Existen diversas aproximaciones para calcular los errores estándar robustos, tanto para variables continuas como discretas. La ventaja del error estándar robusto es que no es necesario estimar el coeficiente de correlación intraclase (CCI) por separado antes de realizar el análisis. La desventaja es que puede ser menos eficiente y, por lo tanto, necesitaría una mayor cantidad de conglomerados que el uso del error estándar dependiente del modelo. En general, los métodos que ajustan el error estándar dan una estimación del efecto del tratamiento que pondera por el tamaño del clúster, de forma que los grupos más grandes tienen mayor peso. (35)

La formulación más conocida de dichos errores es la conocida como error estándar de Huber-White, que ajusta el cálculo usando la variabilidad empírica de los residuos del modelo. (36) Esta sería su formulación:

$$\sigma_{H-W}^2 = \frac{\sum_i^n x_i^2 \varepsilon_i^2}{(\sum_i^n x_i^2)^2}$$

Donde ε son los residuos observados en la propia muestra.

Los errores estándar robustos pueden brindar una mejor evaluación de la variabilidad de las estimaciones de clúster a clúster cuando se violan los supuestos del modelo estadístico, en concreto el supuesto de homocedasticidad.

Otra alternativa para abordar el problema de la no independencia de las observaciones con su correspondiente falta de homogeneidad de la varianza es la utilización de los Modelos Lineales Generalizados (MLG, *Generalized Linear Models*, GLM en la literatura internacional) que utilizan para su estimación métodos de máxima verosimilitud, lo que permite obtener resultados sin tener que suavizar la variable dependiente, con lo que el problema de la heterocedasticidad queda relegado. El MLG puede describirse de esta manera:

$$g\{E(y)\} = x_{ij}\beta_j \quad y \sim F$$

Donde y es la variable dependiente, $g(.)$ es la denominada función de enlace (“link function”), F es la familia de la distribución, x_{ij} representa la matriz de covariables y β_j el vector de los coeficientes. La función de enlace relaciona la media condicional a las covariables. La familia de la distribución F se utiliza para especificar la relación entre la varianza y la media. (37)

Otra forma de abordar el análisis de los ensayos clínicos por clústeres es mediante el método de la ecuación de estimación generalizada (*Generalized Estimating Equations*, GEE). Para su cálculo utiliza un algoritmo de reponderación iterativo para los parámetros y un método robusto para el error estándar. Se podría formular de la siguiente manera:

$$g(E[y_{ij}|x_{ij}]) = x'_{ij} \beta$$

X_{ij} es un vector de covariables

β es el vector de los coeficientes

$g(.)$ es la función de enlace

Y_{ij} es el resultado para el sujeto i del grupo j

Dependiendo de la función de enlace utilizada pueden abordarse datos de diferente naturaleza:

- $g(a)=a$ (función identidad)
- $g(a)=\log(a)$ (para datos cuantitativos, suavizados)
- $g(a)=\log(a/(1-a))$ (logit para datos binarios)

Estos modelos corrigen la dependencia entre los individuos de un mismo grupo asumiendo a priori una cierta estructura de correlación para la variable dependiente medida en cada grupo. Además, son poco exigentes con la distribución de la variable respuesta y ofrecen errores estándar robustos de los coeficientes, que son estables incluso si la correlación elegida a priori es incorrecta o si la intensidad de la correlación entre las observaciones varía entre los diferentes grupos. (38) En su interpretación, sin embargo, no permiten medir la influencia que una característica tiene sobre cada individuo, sino que caracterizan el promedio de respuesta poblacional cuando cambian las variables independientes en el conjunto de la población. Se ha recomendado no utilizar modelos GEE cuando se tienen menos de 20 clústeres, pues los residuos observados que se utilizan en la estimación robusta son estimaciones sesgadas del error de la variable. (34)

Una forma alternativa de abordaje es el análisis mediante modelos multinivel. Los llamados “modelos jerárquicos” o “modelos multinivel” de manera genérica permiten establecer asociaciones entre características individuales y de grupo y un resultado único, por lo que son apropiados para el análisis de los ensayos por clústeres. Pueden incluir variables independientes de nivel individual y de nivel grupal (de clúster), sin incumplir ningún supuesto de aplicación pues se formulan en dos etapas (para el caso más sencillo). En la primera etapa, se define una regresión de nivel individual separada para cada grupo. En una segunda etapa, cada uno de los coeficientes de regresión específicos del grupo o contexto definidos en la ecuación se modelan en función del nivel de grupo. La formulación de este tipo de modelos cuando evaluamos relaciones lineales puede expresarse así: (39)

$$Y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \beta_2 Z_j + \mu_{0j} + \mu_{1j} x_{ij} + \varepsilon_{ij}$$

X_{ij} representa las variables de cada individuo “i” de cada grupo “j”

Z_j es el conjunto de variables contextuales (comunes para el grupo) de cada grupo “j”

μ_{0j} es el efecto aleatorio de la media de cada grupo

μ_{1j} es el efecto aleatorio de la pendiente de cada grupo “j”

ε_{ij} es el efecto aleatorio de un individuo “i” del grupo “j”

Este tipo de modelos permite estudiar la influencia del grupo en la variabilidad de la respuesta una vez que se ha extraído el efecto de variables individuales y de grupo. A su vez permite saber si las medias son diferentes en cada grupo y si la asociación entre la variable independiente y la dependiente es igual o no en cada grupo.

Finalmente señalaremos la posibilidad de la utilización del análisis bayesiano jerárquico de los ensayos por conglomerados, un enfoque flexible e intuitivo para la inferencia estadística. Como en el caso de los ensayos con aleatorización individual, el análisis bayesiano facilita una interpretación probabilística más natural de los resultados y se aleja de las pruebas de hipótesis frecuentistas y los valores de p . (40)

Otro aspecto a considerar en el análisis de los ensayos por clústeres es el manejo de los datos perdidos. Si se decide hacer imputación de datos perdidos existen las mismas posibilidades metodológicas que en otros diseños. Las estrategias de imputación única, como la última observación realizada (*Last Observation Carried Forward*, LOCF) utilizada en estudios longitudinales, o la sustitución por el valor medio de la variable, que reemplaza los datos faltantes con un solo número, reducen la incertidumbre, pero hacen suposiciones poco sólidas ya sea sobre la trayectoria del sujeto o sobre sus características, que pueden llevar a subestimar o sobreestimar las asociaciones. (41)

Existen varias perspectivas a la hora de valorar los mejores métodos de imputación de datos perdidos. Se considera que los datos son perdidos al azar (*missing completely at random*, MCAR) si la pérdida es independiente de los resultados observados y las covariables. Esta suposición es poco probable. Parece más razonable pensar que las pérdidas no se producen de forma aleatoria, pero que pueden explicarse completamente mediante variables en las que hay información completa (*missing at random*, MAR). O se puede asumir que el valor de la variable que falta está relacionado con la razón por la que se ha perdido (*missing not at random*, MNAR). Bajo el supuesto MAR, los métodos de imputación múltiple consideran la incertidumbre completando los datos faltantes de una distribución de valores probables. El problema en este caso es que la mayoría de los procedimientos de imputación múltiple ignoran la estructura multinivel de los ensayos por clústeres, sin incorporar la falta de independencia entre los valores de los sujetos que se encuentran dentro de los grupos. Sí serán válidos los modelos de imputación jerárquicos, si son correctamente especificados. Bajo la perspectiva del MCAR, los modelos de imputación pueden hacerse mediante GEE si hay suficiente número de clústeres. (42)

La interpretación de los resultados de los ensayos por conglomerados puede ser más complicada que la de los ensayos clásicos, ya que las conclusiones pueden relacionarse con los conglomerados, con los participantes en esos conglomerados o con ambos. En ocasiones solo permite valorar estrategias en su conjunto, pero esto tiene utilidad para la toma de decisiones. No siempre se puede individualizar el efecto de un solo agente o de una parte de la intervención, ni sobre un solo aspecto del resultado.

6. Intervenciones complejas

Las intervenciones complejas se describen generalmente como aquellas que contienen distintos componentes que interactúan entre sí, pero hay que tener en cuenta que existen distintos niveles de complejidad y que hay otros elementos que pueden interactuar con la intervención en los que a priori no pensaríamos. (43) Aunque no existe un método ideal para la evaluación de las intervenciones complejas, ya que se pueden utilizar una variedad de métodos enfocados a abordar varias dimensiones de la intervención (evaluación del efecto de la intervención, de su implementación, de los mecanismos de impacto y de los factores contextuales), los ensayos por conglomerados son una de las opciones más utilizadas. (44,45) Las principales dificultades en la evaluación de la intervención compleja se relacionan con: a) la estandarización en el diseño y ejecución de las intervenciones, b) las dificultades logísticas y metodológicas de aplicar diseños experimentales, y c) la complejidad de establecer la relación de causalidad entre la intervención y un determinado resultado. (46)

Las intervenciones complejas son estrategias que pueden incluir diversas tecnologías (fármacos, pruebas diagnósticas, consejo clínico u otro tipo de tratamientos), y lo que interesa al estudiarlas es conocer cómo funciona la estrategia en su conjunto, en lugar de aislar la “eficacia” teórica de cada uno de sus componentes. La complejidad viene definida por la interacción entre sus componentes, por los comportamientos de los que aplican la intervención o de los que la reciben, por los diferentes grupos o niveles de organización a cargo de la intervención y la interdependencia del contexto en el que se está realizando. Si retomamos algunos de los ensayos que hemos comentado previamente podemos identificar fácilmente estos elementos. Por ejemplo, en el estudio PROLACT la intervención está estructurada en sesiones grupales, el grupo condiciona el desarrollo de las sesiones, y puede hacer que varíen, lo que implica distintas técnicas que se adaptan a cada sujeto y a su evolución. En el MULTIPAP la interacción entre profesional de medicina-paciente es un elemento fundamental del resultado terapéutico.

Para su diseño en el año 2000, el Medical Research Council publicó una primera propuesta sobre cómo deben desarrollarse y evaluarse las intervenciones complejas. (47) Esta guía ha sido posteriormente actualizada en 2008 (32) y en el 2019. (33) La evaluación de las intervenciones complejas a través del ensayo aleatorizado requiere un enfoque global en varios pasos que comprende el desarrollo de una intervención desde su marco teórico. En la guía publicada, se propone una evaluación secuencial en cinco etapas que incluyen una fase preclínica (o teórica) y cuatro fases clínicas (I-Modelo, II-Ensayo exploratorio, III-Ensayo Definitivo y IV-Implementación a largo término). (48)

Muchas de las intervenciones a testar constan de varios componentes que, combinados, producen el resultado deseado. Es fundamental describir con detalle las intervenciones a testar, tanto las farmacológicas como las no farmacológicas, requiriendo con frecuencia largas explicaciones que son difíciles de estandarizar. Para facilitar su descripción se ha propuesto elaborar una imagen única que presenta los componentes de todas las intervenciones en el ensayo y compara diferentes brazos de tratamiento (Figura 1). El objetivo es aclarar la estructura de las intervenciones contrastadas y así ayudar a la interpretación de los resultados del ensayo. (49,50)

Figura 1. Representación gráfica de la intervención compleja del ensayo MULTIPAP.

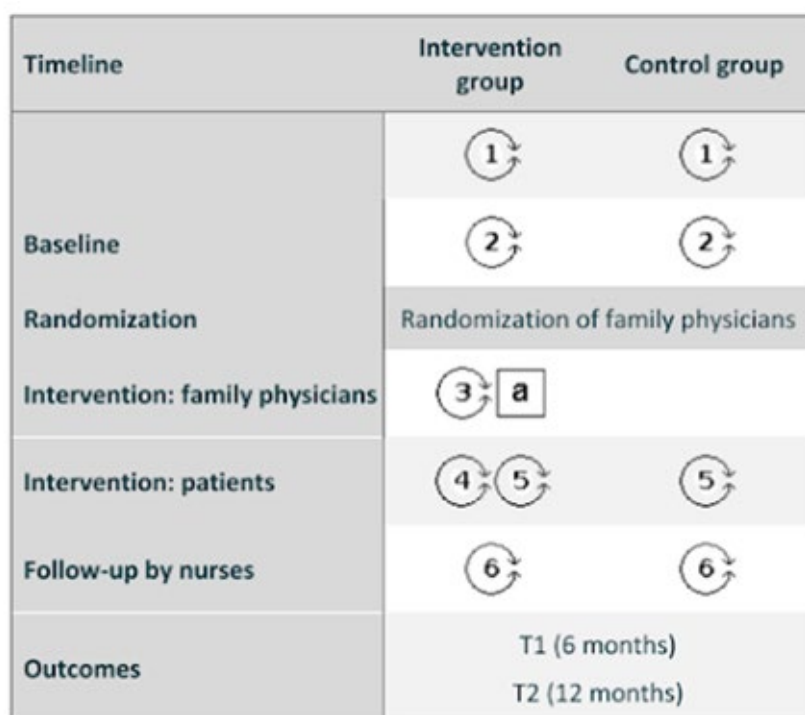


Figura 1. Representación gráfica de la intervención compleja del ensayo MULTIPAP (continuación).

①	Welcome and training session conducted to the 80 family physicians who participate in the study. Presentation of the project. Training in good practices in research and electronic data capture system (3 hours). This session will be certified and uniformly taught in each participating Autonomous Community.
②	Recruitment of patients. Informed consent. Baseline visit (T0). Before starting the intervention, family physicians will collect all baseline variables: socio-demographics, morbidity, pharmacological treatment plan and QoL.
③	Intervention on family physicians (4 weeks, 5 hours per week). Methodology: active and participatory. - Content (I): multimorbidity and polypharmacy. - Content (II): prescription adequacy and adherence. - Content (III): Ariadne principles and their application to clinical practice.
a	Intervention package and support material. Family physicians will receive written documents and a website link to the MOOC courses (slides, videos, tutorials...).
④	Intervention on patients: structured physician-patient interview. It will comprise a structured review of the treatment plan, inclusion of patient preferences, and a pharmacological treatment plan. Visits 1 and 2.
⑤	Usual care: clinical follow-up according to available protocols applicable to the patient and the health service standardized portfolio.
⑥	Follow-up visit by nurses: both groups will receive the usual care plan by their corresponding nurses.

Tomado de: Prados-Torres A, Del Cura González I et al (12).

7. Aspectos éticos en los ensayos por conglomerados

Los ensayos por conglomerados plantean nuevas cuestiones éticas, (8) tanto los ensayos en paralelo como los más novedosos SW-RCT. (51,52,53,54) Una de las cuestiones más debatidas es quién es el sujeto de estudio en los ensayos por conglomerados (55) y, por lo tanto, quién debe firmar el consentimiento informado. (56,57) Si comparamos con los ensayos con aleatorización individual, se ha encontrado que existe una gran variabilidad en este punto en los ensayos aleatorizados por conglomerados, (58) con una mayor probabilidad de información inadecuada de los procedimientos de consentimiento y el uso inapropiado de exenciones al consentimiento. (59)

Algunos autores argumentan que la aleatorización de grupos es una razón para no solicitar el consentimiento informado de los participantes en la investigación. (60) La Declaración de Ottawa sobre el diseño ético y la realización de ensayos

por conglomerados proporcionó a los investigadores y a los comités de ética de la investigación una orientación inicial sobre este tema. (61) Se define a un participante en la investigación como cualquier individuo cuyos intereses pueden verse directamente afectados por los procedimientos de investigación: cualquier persona que recibe una intervención en el contexto de investigación con la que se interactúa para la recogida de datos o cuyos datos privados se utilizan. En los ensayos por conglomerados es la unidad de intervención y no la unidad de aleatorización la que marca las cuestiones relativas al consentimiento informado. (62) Tanto los profesionales sanitarios como los pacientes pueden ser participantes en la investigación en estos ensayos, algo a considerar también al recoger el consentimiento. Por ejemplo, si el médico es el destinatario de la intervención del estudio y el paciente solo tiene que recoger su información sanitaria identificable, los investigadores deben solicitar el consentimiento para la intervención del estudio por parte del médico y el consentimiento para la recogida de datos del paciente.

Se contemplan algunas situaciones que pueden ser objeto de una exención del consentimiento si se cumplen las siguientes condiciones:

1. La investigación no sería factible sin dicha exención o una alteración del consentimiento,
2. tiene un valor social importante y
3. las intervenciones del estudio y los procedimientos de recogida de datos no suponen más que un riesgo mínimo.

En un ensayo por conglomerados la obtención del consentimiento puede ser inviable cuando los individuos no pueden negarse a participar (ver los dos primeros ejemplos de la tabla 1). El criterio de inviabilidad es esencial porque la necesidad del consentimiento informado no se refiere únicamente, ni siquiera principalmente, al riesgo. El riesgo mínimo se refiere a los riesgos de la vida cotidiana para personas normales y sanas, incluidos los riesgos de los exámenes físicos rutinarios o la revisión de historiales médicos.

La guía publicada en 2021 por Nix *et al.* (62) propone un marco de tres pasos para reflexionar sobre el consentimiento informado en los ensayos por conglomerados:

- a) Identificar a los participantes en la investigación: ¿Quién es el sujeto de estudio?,
- b) identificar el elemento o elementos del estudio a los que están expuestos los participantes en la investigación, y
- c) determinar si es apropiada una exención del consentimiento para cada elemento del estudio.

La necesidad de obtención del consentimiento informado de los individuos del conglomerado dependerá de si la intervención se aplica a nivel del grupo o del individuo. Una recomendación útil es “Consiga el consentimiento siempre que pueda”. Podemos identificar dos tipos de intervenciones en el contexto de un ensayo en conglomerados:

- a) En el ensayo clínico aleatorizado por grupos-grupo: la intervención del estudio se realiza a nivel del conglomerado o grupo, por lo que puede ser muy difícil para los miembros del grupo evitar la intervención si ellos no desean participar en el ensayo. Las intervenciones a nivel de clúster se aplican a la comunidad, el hospital o el grupo social en su conjunto y no pueden ser evitadas por los miembros individuales del clúster. Cuando los miembros del grupo no pueden evitar la exposición a la intervención, la denegación del consentimiento carece de sentido. El uso de una renuncia al consentimiento para las intervenciones a nivel de conglomerado es apropiado siempre que la intervención suponga un riesgo mínimo. Cuando la recogida de datos se produce a nivel individual, se requiere el consentimiento informado para la recogida de datos. Se contempla la figura de “*gatekeeper*” o “guardianes”, (63) personas u organismos que representan los intereses de los miembros de las agrupaciones o las organizaciones, y que pueden tener un papel cuando hay dificultades para obtener el consentimiento informado debido a la aleatorización de los grupos, las intervenciones a nivel de grupo y el tamaño de los grupos, pero existe debate sobre cuál debe ser su función y los límites de su autoridad. En general, los guardianes no tienen autoridad para dar el consentimiento por delegación a los miembros del grupo, pero pueden proteger los intereses del grupo al que representan, como hospitales, residencias de ancianos, centros de salud y escuelas. En estos entornos gestionan también el permiso de acceso a la organización.
- b) En el ensayo clínico aleatorizado por grupos-individuo: la intervención está centrada en el individuo dentro del grupo, dirigiéndose a los miembros individuales del grupo. En estos casos se debe solicitar el consentimiento tanto a nivel de grupo como de participante individual antes de la aleatorización. Una vez que los grupos han sido asignados aleatoriamente, ya no se puede pedir a los participantes en los grupos su consentimiento para ser asignados aleatoriamente a recibir cualquiera de las intervenciones, sino solo su consentimiento para recibir la intervención a la que se ha asignado su grupo y para el seguimiento. Se busca así evitar la posibilidad de un sesgo de selección posterior a la aleatorización.

Las intervenciones a nivel profesional se realizan a profesionales sanitarios y, por tanto, son participantes en la investigación. Cuando los profesionales de la salud son participantes en la investigación, debe obtenerse su consentimiento informado, a menos que se cumplan las condiciones para una exención del consentimiento. Los pacientes pueden no ser participantes de la investigación en los ensayos que evalúan intervenciones a nivel profesional, a menos que sean el destinatario de la intervención del estudio, se interactúe con ellos para la recogida de datos o se utilice su información privada identificable. En estos casos es necesario el consentimiento para la recogida de datos.

En las intervenciones a nivel individual, las consideraciones sobre el consentimiento informado son similares en los ensayos con aleatorización de individuos y en los aleatorios por conglomerados con intervenciones a nivel individual, ya que prueban los mismos tipos de intervenciones. Si una intervención a nivel individual no puede ser objeto de una exención de consentimiento en un ensayo clásico, la intervención no debería recibir una exención de consentimiento en un ensayo por conglomerados. Si se solicita el consentimiento para una intervención en la práctica clínica, como en el caso de un fármaco o una vacuna, la exención del consentimiento nunca es apropiada para esa intervención en un ensayo.

8. Como reportar con calidad un ensayo por conglomerados

La declaración CONSORT (*Consolidated Standards of Reporting Trials*), una directriz diseñada para mejorar la transparencia y la calidad de los informes de los ensayos controlados aleatorios, cuenta con varias ampliaciones, entre ellas las de los ensayos por conglomerados, publicada en 2004 (64) y actualizada en 2010. (65) A pesar de ello, el cumplimiento de las directrices metodológicas y de presentación de informes para los ensayos por conglomerados sigue siendo subóptimo y se necesitan más esfuerzos para mejorar tanto la presentación de informes como la metodología.

En la revisión sistemática de Chan CL *et al.*, (66) en 2017, destacaba cómo en la mayoría de los ensayos estaba definido el conglomerado (100 %) y declarado el número de conglomerados aleatorizados (94 %), pero pocos informaron de cómo el diseño del conglomerado afectó al tamaño de la muestra (17 %) o si se solicitó el consentimiento de los conglomerados (11 %). Incluso con la extensión del CONSORT para la aleatorización por grupos, la notificación de los elementos del tama-

ño de la muestra específicos de estos ensayos sigue siendo un déficit importante. (67,68) Es clave también revisar la propuesta de la extensión CONSORT *cluster* para elaborar los flujogramas de los ensayos por conglomerados y proporcionar la información adecuada de los distintos niveles.

Se cuenta también con extensiones del CONSORT para los ensayos por conglomerados SW-RCT. (69,70,71) Una mención específica refiere las recomendaciones de calidad para informar las intervenciones complejas TIDieR. (72) Los 12 ítems resultantes de la lista TIDieR (nombre abreviado, por qué, qué [materiales], qué [procedimiento], quién realizó la intervención, cómo, dónde, cuándo y cuánto, adaptaciones, modificaciones, cómo de bien [planificado], cómo de bien [observado]) (73) son una ampliación de la declaración CONSORT 2010 (ítem 5) y de la declaración SPIRIT 2013 (ítem 11).

9. Referencias

1. Campbell MK, Piaggio G, Elbourne DR, *et al.* CONSORT Group. CONSORT 2010 statement: extension to cluster randomised trials. *BMJ* 2012;345: e5661.
2. BMJ Education and Debate Trials randomised in clusters. *BMJ* 1997; 315:600.
3. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. *Stat Med* 2007; 26:2-19.
4. Lorenz E, Köpke S, Pfaff H, Blettner M. Cluster-Randomized Studies. *Dtsch Arztebl Int.* 2018;115(10):163-168.
5. Pletcher MJ, Flaherman V, Najafi N, Patel S, Rushakoff RJ, Hoffman A, *et al.* Randomized Controlled Trials of Electronic Health Record Interventions: Design, Conduct, and Reporting Considerations. *Ann Intern Med.* 2020 Jun 2;172(11 Suppl): S85-S913.
6. Dron L, Taljaard M, Cheung YB, Grais R, Ford N, Thorlund K, *et al.* The role and challenges of cluster randomised trials for global health. *Lancet Glob Health.* 2021 May;9(5): e701-e710. Erratum in: *Lancet Glob Health.* 2022 Feb;10(2): e185.

7. Samsa GP, Winger JG, Cox CE, Olsen MK. Two Questions About the Design of Cluster Randomized Trials: A Tutorial. *J Pain Symptom Manage.* 2021 Apr;61(4):858-863.
8. Prieto Martín de los Santos E, Portolés Pérez A. Aspectos metodológicos y éticos de los ensayos clínicos en conglomerados. En: Dal-Ré R, Carné X, Gracia D (ed). *Luces y sombras en la investigación clínica*, Madrid Triacastela 2013 pp 217-240.
9. Chalkidou K, Tunis, S, Whicher D, Fowler R, Zwarenstein, M. The role for pragmatic randomized controlled trials (pRCTs) in comparative effectiveness research. *Clinical trials.* 2012; 9(4): 436-446.
10. Black N, Burke L, Forrest CB, Sieberer UH, Ahmed S, Valderas JM, Bartlett SJ, Alonso J. Patient-reported outcomes: pathways to better health, better services, and better societies. *Qual Life Res.* 2016 ;25(5):1103-12.
11. Norton WE, Loudon K, Chambers DA, Zwarenstein M. Designing provider-focused implementation trials with purpose and intent: introducing the PRECIS-2-PS tool. *Implement Sci.* 2021;16(1):7.
12. Prados-Torres A, Del Cura-González I, Prados-Torres D, López-Rodríguez JA, Leiva-Fernández F, Calderón-Larrañaga A, *et al.* Effectiveness of an intervention for improving drug prescription in primary care patients with multimorbidity and polypharmacy: study protocol of a cluster randomized clinical trial (Multi-PAP project). *Implement Sci.* 2017 Apr 27;12(1): 54.
13. Cura-González, I.; López-Rodríguez, J.A.; Leiva-Fernández, F.; Gimeno-Miguel, A.; Poblador-Plou, B.; López-Verde, F. *et al.* How to Improve Healthcare for Patients with Multimorbidity and Polypharmacy in Primary Care: A Pragmatic Cluster-Randomized Clinical Trial of the MULTIPAP Intervention. *J. Pers. Med.* 2022; 12, 752.
14. Leiva-Fernández F, Prados-Torres JD, Prados-Torres A, Del-Cura-González I, Castillo-Jimena M, López-Rodríguez JA *et al.* Training primary care professionals in multimorbidity management: Educational assessment of the eMULTIPAP course. *Mech Ageing Dev.* 2020; 192:111354.
15. Rico-Blázquez M, García-Sanz P, Martín-Martín M, López-Rodríguez JA, Morey-Montalvo M, Sanz-Cuesta T, [...] Cura-González I Effectiveness of a home-based nursing support and cognitive restructuring intervention on the quality of life of family caregivers in primary care: A pragmatic cluster-randomized controlled trial. *Int J Nurs Stud.* 2021; 120:103955.
16. Santamaría-Martín MJ, Martín-Iglesias S, Schwarz C, Rico-Blázquez M, [...] Del-Cura-González I. Effectiveness of a group educational interven-

- tion - prolact - in primary care to promote exclusive breastfeeding: a cluster randomized clinical trial. *BMC Pregnancy Childbirth*. 2022;22(1):132.
17. Escortell-Mayor E, Del Cura-González I, Ojeda-Ruiz E, Sanz-Cuesta T, *et al*. A primary healthcare information intervention for communicating cardiovascular risk to patients with poorly controlled hypertension: The Education and Coronary Risk Evaluation (Educore) study-A pragmatic, cluster-randomized trial. *PLoS One*. 2020;15(1):e0226398.
 18. Minué-Lorenzo C, Olano-Espinosa E, Del Cura-González I, Vizcaíno-Sánchez JM *et al*. Subsidized pharmacological treatment for smoking cessation by the Spanish public health system: A randomized, pragmatic, clinical trial by clusters. *Tob Induc Dis*. 2019; 17:64.
 19. Bolzern J, Mnyama N, Bosanquet K, Torgerson DJ. A review of cluster randomized trials found statistical evidence of selection bias. *J Clin Epidemiol*. 2018; 99:106-112.
 20. Grayling MJ, Wason JM, Mander AP. Stepped wedge cluster randomized controlled trial designs: a review of reporting quality and design features. *Trials*. 2017;18(1):33.
 21. Hemming K, Taljaard M, Grimshaw J. Introducing the new CONSORT extension for stepped-wedge cluster randomised trials. *Trials*. 2019;20(1):68.
 22. Martin J, Taljaard M, Girling A, Hemming K. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ Open*. 2016;6(2): e010166.
 23. Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. John Wiley & Sons, 2012.
 24. Hemming K, Taljaard M, McKenzie JE, *et al*. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ*. 2018;363: k1614.
 25. Hargreaves JR, Copas AJ, Beard E, *et al*. Five questions to consider before conducting a stepped wedge trial. *Trials*. 2015; 16:350.
 26. Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? *Int J Epidemiol*. 2020;49(3):1043-1052.
 27. Kennedy-Shaffer L, Hughes MD. Sample size estimation for stratified individual and cluster randomized trials with binary outcomes. *Stat Med*. 2020;39(10):1489-1513.

28. Hemming K, Girling AJ, Sitch AJ, *et al.* Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Med Res Methodol* 2011; 11:102-11.
29. Rutterford C, Taljaard M, Dixon S, *et al.* Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol* 2015; 68:716-23.
30. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol*. 2004;57: 785-794.
31. Wang J, Cao J, Zhang S, Ahn C. Sample size determination for stepped wedge cluster randomized trials in pragmatic settings. *Stat Methods Med Res*. 2021 Jun 17:9622802211022392.
32. Turner RM, White IR, Croudace T, PIP Study Group. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Stat Med* 2007; 26:274-89.
33. Offorha BC, Walters SJ, Jacques RM. Statistical analysis of publicly funded cluster randomised controlled trials: a review of the National Institute for Health Research Journals Library. *Trials*. 2022;23(1):115.
34. Williams, R. L.). A note on robust variance estimation for cluster-correlated data. *Biometrics* 2000; 56(2), 645-646.
35. Campbell, M. J., Donner, A., & Klar, N. Developments in cluster randomized trials and statistics in medicine. *Statistics in medicine*, 2007; 26(1), 2-19.
36. Mansournia MA, Nazemipour M, Naimi AI, Collins GS, Campbell MJ. Reflection on modern methods: demystifying robust standard errors for epidemiologists. *Int J Epidemiol*. 2021;50(1):346-351.
37. Hardin, J., & Hilbe, J. (2018). *Generalized Linear Models and Extensions* (Fourth Edit). College Station: Stata Press.
38. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med*. 1998;17(11):1261-91.
39. Diez Roux AV. A glossary for multilevel analysis. *J Epidemiol Community Health*. 2002;56(8):588-594.
40. Jones BG, Streeter AJ, Baker A, Moyeed R, Creanor S. Bayesian statistics in the design and analysis of cluster randomised controlled trials and their reporting quality: a methodological systematic review. *Syst Rev*. 2021 Mar 31;10(1):91.

41. Kenward MG, Molenberghs G. Last observation carried forward: a crystal ball? *J Biopharm Stat.* 2009; 19:872-88.
42. Fiero MH, Huang S, Oren E, Bell ML. (). Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials.* 2016; 17(1), 1-10.
43. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ.* 2008;337: a1655.
44. Oakley A, Strange V, Bonell C, Allen E, Stephenson J, RIPPLE Study Team. Process evaluation in randomised controlled trials of complex interventions. *BMJ* 2006;332(7538):413-6.
45. Minary L, Trompette J, Kivits J, Cambon L, Tarquinio C, Alla F. Which design to evaluate complex interventions? Toward a methodological framework through a systematic review. *BMC Med Res Methodol.* 2019;19(1):1-9.
46. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. MRC Developing and evaluating complex interventions. 2019;1-39. Available from: www.mrc.ac.uk/complexinterventionsguidanc
47. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, *et al.* Framework for design and evaluation of complex interventions to improve health Framework for trials of complex interventions. *Br Med J.* 2000;321(7262):694-6.
48. Medical Research Council. A framework for development and evaluation of RCTs for Complex Interventions to improve health [Internet]. London; 2000 [cited 2022 May 15]. Available from: <https://mrc.ukri.org/documents/pdf/rcts-for-complex-interventions-to-improve-health/>
49. Perera R, Heneghan C, Yudkin P. Graphical method for depicting randomised trials of complex interventions. *BMJ.* 2007;334(7585):127-129.
50. Hooper R, Froud RJ, Bremner SA, Perera R, Eldridge S. Cascade diagrams for depicting complex interventions in randomised trials. *BMJ.* 2013; 15;347: f6681.
51. Taljaard M, Hemming K, Shah L, Giraudeau B, Grimshaw JM, Weijer C. Inadequacy of ethical conduct and reporting of stepped wedge cluster randomized trials: Results from a systematic review. *Clin Trials.* 2017;14(4):333-341.
52. Aspectos éticos en los ensayos por cluster
53. Weijer C, Grimshaw JM, Taljaard M, Binik A, Boruch R, Brehaut JC, Donner A, Eccles MP, Gallo A, McRae AD, Saginur R, Zwarenstein M.

- Ethical issues posed by cluster randomized trials in health research. *Trials*. 2011 Apr 20; 12:100.
54. Sim J, Dawson A. Informed consent and cluster-randomized trials. *Am J Public Health*. 2012 ;102(3):480-5.
 55. McRae A, Weijer C, Binik A, *et al.* Who is the research subject in cluster randomized trials in health research? *Trials*. 2011; 12: 183.
 56. Nix HP, Weijer C, Brehaut JC, Forster D, Goldstein CE, Taljaard M. Informed consent in cluster randomised trials: a guide for the perplexed. *BMJ Open*. 2021;11(9):e054213.
 57. Giraudeau B, Caille A, Le Gouge A, Ravaud P. Participant informed consent in cluster randomized trials: review. *PLoS One*. 2012;7(7):e40436..
 58. Taljaard M, Brehaut JC, Weijer C, Boruch R, Donner A, Eccles MP, McRae AD, Saginur R, Zwarenstein M, Grimshaw JM. Variability in research ethics review of cluster randomized trials: a scenario-based survey in three countries. *Trials*. 2014; 15:48.
 59. McRae A, Taljaard M, Weijer C, Bennett C, Skea Z, Boruch R, Brehaut J, Eccles M, Grimshaw J, Donner A. Reporting of patient consent in health-care cluster randomised trials is associated with the type of study interventions and publication characteristics. *J Med Ethics*. 2013;39(2):119-249.
 60. McRae AD, Weijer C, Binik A, Grimshaw JM, *et al.* When is informed consent required in cluster randomized trials in health research? *Trials*. 2011; 12:202.
 61. Weijer C, Grimshaw JM, Eccles MP, *et al.* The Ottawa statement on the ethical design and conduct of cluster randomized trials. *PLoS Med* 2012;9: e1001346.
 62. Nix HP, Weijer C, Brehaut JC, Forster D, *et al.* Informed consent in cluster randomised trials: a guide for the perplexed. *BMJ Open*. 2021 ;11(9): e054213.
 63. Gallo, A., Weijer, C., White, A. *et al.* What is the role and authority of gatekeepers in cluster randomized trials in health research? *Trials* 2012; 13, 116.
 64. Campbell MJ. Extending CONSORT to include cluster trials. *BMJ*. 2004;328(7441):654-5.
 65. Campbell MK, Piaggio G, Elbourne DR, *et al.* CONSORT Group. CONSORT 2010 statement: extension to cluster randomised trials. *BMJ* 2012;345: e5661

66. Chan CL, Leyrat C, Eldridge SM. Quality of reporting of pilot and feasibility cluster randomised trials: a systematic review. *BMJ Open*. 2017;7(11): e016970.
67. Ivers NM, Taljaard M, Dixon S, Bennett C, *et al*. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ*. 2011;343: d5886.
68. Rutterford C, Taljaard M, Dixon S, *et al*. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol* 2015; 68:716-23.
69. Davey C, Hargreaves J, Thompson JA, Copas AJ, *et al*. Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. *Trials*. 2015; 16:358.
70. Hemming K, Taljaard M, McKenzie JE, Hooper R, *et al*. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ*. 2018 Nov 9;363: k1614.
71. Hemming K, Taljaard M, Grimshaw J. Introducing the new CONSORT extension for stepped-wedge cluster randomised trials. *Trials*. 2019 Jan 18;20(1):68.
72. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, *et al*. better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014 Mar 7;348: g1687.
73. <https://www.equator-network.org/wp-content/uploads/2014/03/Guia-y-lista-de-verificacion-TiDier-Traduccion-espanol.pdf>

Capítulo 13.

Estudios diagnósticos.

Validez y fiabilidad

IÑAKI IMAZ IGLESIA, ESTHER ELENA GARCÍA CARPINTERO
y JESÚS GONZÁLEZ ENRÍQUEZ

1. Introducción

2. Validez o exactitud de una prueba diagnóstica

- 2.1. La influencia de la prevalencia en los resultados de una prueba diagnóstica
- 2.2. Las razones de máxima verosimilitud
- 2.3. El umbral de positividad de una prueba. La curva de características operativas (curva ROC)

3. Fiabilidad o concordancia de una prueba diagnóstica

- 3.1. El coeficiente de concordancia Kappa
- 3.2. El Coeficiente de Correlación Intraclass

4. Lectura crítica de estudios diagnósticos

5. Referencias

1. Introducción

El proceso diagnóstico recoge información del estado de salud del paciente, usando la información que proporcionan sus características personales, los síntomas, la historia clínica, el examen físico, las pruebas de laboratorio y otras pruebas adicionales. Los objetivos del uso de pruebas diagnósticas son múltiples: aumentar la certeza de la presencia o ausencia de la enfermedad o de alteraciones fisiológicas, identificar marcadores de riesgo, apoyar el manejo clínico, monitorizar el curso clínico o la respuesta al tratamiento, evaluar el pronóstico de la enfermedad o medir las aptitudes físicas y mentales (1).

Una prueba diagnóstica es cualquier enfoque utilizado para reunir información con el fin de tomar una decisión clínica. Desde el punto de vista estadístico, hay que tener en cuenta que el proceso de toma de decisiones se basa en la probabilidad y que el objetivo de la prueba diagnóstica es desplazar la probabilidad estimada de la enfermedad o condición en estudio hacia cualquiera de los extremos de la escala de probabilidad, hacia el 0 donde se descarta o hacia el 1 donde se confirma.

Una buena prueba diagnóstica es aquella que es capaz de discriminar o clasificar correctamente a los miembros de la población en que se aplica, ofreciendo un resultado positivo en los enfermos o portadores de la condición buscada y negativo en los sanos o no portadores. La capacidad de la prueba de diagnosticar adecuadamente se evalúa fundamentalmente estudiando su exactitud o validez y su fiabilidad o concordancia.

La exactitud no es una propiedad fija de la prueba diagnóstica, sino que varía según las características de la población y el entorno en que se aplique, y puede estar condicionada por los resultados de pruebas anteriores. La valoración de la utilidad de una prueba diagnóstica tiene que tener presente el uso potencial destinado para la prueba, ya que este puede ser el de remplazar a una prueba existente, como complementario de una prueba previa (triaje) o como prueba adicional en un protocolo diagnóstico.

El diagnóstico es un proceso multiprueba o multivariable de estimación y actualización de la probabilidad de la presencia de la enfermedad, dadas las combinaciones de resultados de las distintas pruebas diagnósticas realizadas (2). En la práctica clínica es importante conocer qué pruebas diagnósticas son redundantes y cuáles tienen un valor predictivo real. En consecuencia, los estudios de exactitud diagnóstica deben demostrar qué pruebas diagnósticas pueden aumentar o disminuir la probabilidad de presencia de la enfermedad (2).

A continuación, vamos a detallar los principales indicadores a utilizar para la evaluación de una prueba diagnóstica. Estos son la validez o exactitud y la fiabilidad o concordancia. Finalmente, en la última sección se exponen las principales herramientas que facilitan la lectura crítica de estudios diagnósticos.

2. Validez o exactitud de una prueba diagnóstica

La validez es el grado en que una prueba mide lo que se supone debe medir y corresponde a la exactitud diagnóstica. Es la capacidad de la prueba de discriminar entre los casos y los no casos para una enfermedad o condición clínica. La validación de una prueba diagnóstica se realiza mediante la comparación de sus resultados con los obtenidos con el mejor instrumento de medida del fenómeno estudiado o con el verdadero resultado si éste es conocido, o sea, mediante la comparación del test con un «estándar de oro». Cuanto mayor grado de acuerdo tenga la prueba diagnóstica con el estándar de oro más válida será.

Algunas veces el estándar de oro es una prueba diagnóstica relativamente simple y económica, como por ejemplo el cultivo de garganta para estreptococo β -hemolítico del grupo A para validar la impresión clínica de faringitis estreptocócica. Sin embargo, más frecuentemente debe recurrirse a exámenes complejos, costosos o arriesgados (biopsia, radiología con contraste, autopsia). En ocasiones, no se dispone de ningún estándar de oro, por la naturaleza de la condición a medir o por la ausencia de información científica suficiente. En estos casos, se utilizan criterios diagnósticos diseñados por expertos o resultados de un conjunto de pruebas agrupadas.

El grado de acuerdo entre la prueba diagnóstica y el estándar de oro se representa en una tabla de contingencia, donde en las columnas se representa el resultado del estándar de oro que clasifica a los sujetos en enfermos y sanos y en las filas el resultado de la prueba diagnóstica (positiva o negativa). La tabla de contingencia contiene el recuento de los casos que se dividen en verdaderos positivos (enfermos con prueba diagnóstica positiva), verdaderos negativos (sanos con prueba diagnóstica negativa), falsos positivos (sanos con prueba diagnóstica positiva) y falsos negativos (enfermos con prueba diagnóstica negativa). En la tabla 1 se exponen los principales indicadores para evaluar la validez de una prueba diagnóstica.

Tabla 1. Relación entre una prueba diagnóstica y un estándar de oro

		Enfermedad		
		+	-	
Prueba	+	a VP	b FP	a+b
	-	c FN	d VN	c+d
		a+c	b+d	a+b+c+d

$$\text{Sensibilidad} = \frac{a}{a+c} \quad \text{Especificidad} = \frac{d}{b+d}$$

$$\text{Valor Predictivo Positivo (VPP)} = \frac{a}{a+b}$$

$$\text{Valor Predictivo Negativo (VPN)} = \frac{d}{c+d}$$

$$\text{Prevalencia} = \frac{a+c}{a+b+c+d}$$

- La **sensibilidad** de una prueba es la probabilidad de que un sujeto enfermo sea adecuadamente clasificado. Una prueba será tanto más sensible cuanto menos falsos negativos (FN) proporcione. En las pruebas de cribado generalmente se utilizan test de alta sensibilidad para captar a casi todos los enfermos. También en enfermedades graves tratables, en las que no debe permitirse la pérdida de ningún caso.
- La **especificidad** de una prueba es la probabilidad de que un sujeto sano sea clasificado correctamente. Una prueba será tanto más específica cuanto menos falsos positivos (FP) proporcione. En general, las pruebas de confirmación diagnóstica deben tener una elevada especificidad, así como en la identificación de enfermedades graves no tratables, en las que un resultado falso positivo puede producir serios trastornos al individuo.
- El **valor predictivo positivo** de una prueba diagnóstica es la probabilidad de que un sujeto con un resultado positivo sea un individuo enfermo. El **valor predictivo negativo** es la probabilidad de que un sujeto con un resul-

tado negativo en la prueba sea un individuo sano. Los valores predictivos dependen de la prevalencia de la enfermedad que se quiere diagnosticar.

- La **seguridad** de una prueba diagnóstica se valora mediante el valor predictivo positivo y negativo, ya que nos dice en qué grado los resultados positivos y negativos de la prueba son correctos. Los clínicos necesitan saber, antes de actuar y tratar, con qué seguridad el paciente tiene la enfermedad cuando la prueba diagnóstica es positiva.

Ejemplo

Durante la pandemia producida por el virus SARS-CoV-2 se realizó un gran esfuerzo en desarrollar nuevos test que permitieran un rápido diagnóstico para identificar, tratar y aislar rápidamente a los sujetos con COVID-19. Para el diagnóstico de la COVID-19 se considera como estándar de oro la prueba de la técnica de la reacción en cadena de la polimerasa (PCR). Durante este periodo se desarrollaron test rápidos para el diagnóstico de la COVID-19, como los test de antígenos, que permiten detectar proteínas del SARS-CoV-2. Una revisión sistemática estimó que, en promedio, comparados con una PCR, estos test tenían una sensibilidad del 56 % y una especificidad del 99 % (3). En la tabla 2 se muestran los resultados de validez diagnóstica de estos test en una población con alta prevalencia de infección como podría ser la que acude a un punto diagnóstico por presentar síntomas compatibles con la enfermedad en una zona con alta transmisión comunitaria. La especificidad de la prueba es del 99 %, lo que indica que la prueba clasifica correctamente a los sujetos sanos. Sin embargo, los resultados de sensibilidad indican que un 44 % de los infectados no habrían sido diagnosticados, por lo que este test no sería adecuado para la identificación de casos en ese contexto.

Los valores predictivos positivo y negativo de la prueba fueron del 99 % y 57 % respectivamente. Estos valores indican que, en el caso de un resultado positivo, hay un 99 % de probabilidad de que el sujeto tenga el SARS-CoV-2 y, en el caso de un resultado negativo, se puede descartar la infección con una probabilidad del 57 %.

Tabla 2. Resultados del test de antígenos para el diagnóstico de la COVID-19

		PCR		
		+	-	
Test de antígeno	+	335 VP	2 FP	337
	-	261 FN	345 VN	606
		596	347	943

$$\text{Sensibilidad} = \frac{335}{335 + 261} = 0,56 \quad \text{Especificidad} = \frac{2}{2 + 345} = 0,99$$

$$\text{VPP} = \frac{335}{335 + 2} = 0,99 \quad \text{VPN} = \frac{345}{261 + 345} = 0,57$$

2.1. La influencia de la prevalencia en los resultados de una prueba diagnóstica

Una prueba diagnóstica de calidad debe tener una sensibilidad y especificidad altas (próximas al 100 %), pero para garantizar buenos resultados también deben tenerse en cuenta las condiciones de aplicación de la prueba. La prevalencia de enfermedad en la población en la que se aplica el test, aunque no modifica la sensibilidad y especificidad, puede modificar los valores predictivos positivo y negativo convirtiendo una prueba de calidad en una intervención diagnóstica de poca utilidad.

Una prevalencia alta favorece la identificación de casos, pero dificulta descartarlos con la obtención de un resultado negativo. Por el contrario, una prevalencia baja da más valor al resultado negativo, pero menor al resultado positivo, ya que la probabilidad de estar enfermo es pequeña.

En el ejemplo anterior, el test de antígenos era utilizado en un contexto de alta prevalencia (63,2 %). Si utilizamos ese mismo test en un supuesto de baja prevalencia los resultados son muy diferentes. En la tabla 3 se exponen los resultados en un supuesto de una prevalencia del 5 %, suponiendo una cohorte de 1000 personas, y con la misma sensibilidad (56 %) y especificidad (99 %) de la prueba. En este contexto, el VPP baja del 99 % al 85 %, y el VPN sin embargo se eleva del 57 % al 98 %. Por tanto, en este caso el 85 % de los sujetos con una prueba positiva tendrían la enfermedad, por lo que la utilización del test en este contexto tiene menos valor para confirmar enfermedad, mientras que sí sirve para confirmar que los individuos con prueba negativa están libres de la COVID-19, con una probabilidad del 98 %. En el caso de una población de 1000 sujetos con prevalencia alta de la enfermedad (20 %), los valores de VPP y VPN serían 97 % y 90 % respectivamente (tabla 4). Estos datos indican que, en un escenario de alta prevalencia, el test de antígenos analizado nos proporciona una mayor seguridad en los resultados positivos, con un menor número de falsos positivos.

Tabla 3. Resultados del test de antígenos en una población con baja prevalencia (5 %) de COVID-19

		PCR		
		+	-	
Test de antígeno	+	28 VP	5 FP	33
	-	22 FN	945 VN	967
		50	950	1000

$$\text{Sensibilidad} = \frac{28}{28 + 22} = 0,56 \quad \text{Especificidad} = \frac{945}{945 + 5} = 0,99$$

$$\text{Prevalencia} = \frac{28 + 22}{28 + 22 + 5 + 950} = \frac{50}{1000} = 0,05$$

$$\text{VPP} = \frac{28}{28 + 5} = 0,85 \quad \text{VPN} = \frac{945}{22 + 945} = 0,98$$

Tabla 4. Resultados del test de antígenos en una población con alta prevalencia (20%) de COVID-19

		PCR		
		+	-	
Test de antígeno	+	112 VP	4 FP	116
	-	88 FN	796 VN	884
		200	800	1000

$$\text{Sensibilidad} = \frac{112}{112 + 88} = 0,56 \quad \text{Especificidad} = \frac{796}{796 + 4} = 0,99$$

$$\text{Prevalencia} = \frac{112 + 88}{112 + 4 + 88 + 796} = \frac{200}{1000} = 0,2$$

$$\text{VPP} = \frac{112}{112 + 4} = 0,97 \quad \text{VPN} = \frac{796}{796 + 88} = 0,90$$

2.2. Las razones de máxima verosimilitud

Un indicador útil para discriminar entre pruebas diagnósticas por su valor informativo es la razón de verosimilitud o *likelihood ratio*, que mide cuánto de probable es un resultado según la presencia o ausencia de la enfermedad. Este indicador no tiene un umbral que establezca cuando la prueba diagnóstica tiene poco valor, pero puede ser útil para conocer los resultados que aportará la realización de la prueba. Además, este indicador no depende de la prevalencia de la enfermedad, lo que permite comparar diferentes pruebas para un mismo diagnóstico.

La razón de verosimilitud positiva (RV+) se calcula dividiendo la probabilidad de un resultado positivo en los pacientes enfermos entre la probabilidad de un resultado positivo entre los sanos, es decir, es el cociente entre la fracción de verdaderos positivos (sensibilidad) y la fracción de falsos positivos (1 – especificidad):

$$RV+ = \frac{\text{sensibilidad}}{1 - \text{especificidad}}$$

Cuanto más elevado es el resultado de la fórmula anterior mejor es la prueba. Por ejemplo, se ha estimado que una tomografía computerizada abdominal para el diagnóstico de una lesión esplénica tiene una sensibilidad del 95 % y una especificidad del 96 %, resultando una razón de máxima verosimilitud de un resultado positivo de 23,8. Sin embargo, los termogramas para la detección del cáncer de mama en la mujer tienen una sensibilidad del 61 % y una especificidad del 74 %, dando una razón de máxima verosimilitud de un resultado positivo mucho más baja (2,35).

La representación gráfica del numerador y denominador de la razón de máxima verosimilitud puede servir de ayuda para discriminar cuál es el punto que ofrece mejores resultados. El siguiente apartado se dedica a la curva ROC, gráfico estándar que representa los puntos de equilibrio entre sensibilidad y especificidad.

La razón de verosimilitud negativa (RV-) se calcula dividiendo la probabilidad de un resultado negativo en presencia de la enfermedad entre la probabilidad de un resultado negativo en ausencia de la misma, es decir, es el cociente entre la fracción de falsos negativos (1 – sensibilidad) y la fracción de verdaderos negativos (especificidad).

$$RV- = \frac{1 - \text{sensibilidad}}{\text{especificidad}}$$

En este caso, cuanto menor sea el resultado de la razón de verosimilitud negativa mejor es la prueba diagnóstica.

Una razón de verosimilitud superior a 1 indica que el resultado de la prueba está asociado con la presencia de la enfermedad, mientras que una razón de verosimilitud inferior a 1 está asociado a la ausencia de la enfermedad. Se considera que un valor de $RV+$ mayor de 10 y un $RV-$ de 0,1 indican un cambio relevante en la probabilidad pre-test y tendrían una utilidad clínica muy alta, mientras que una $RV+$ inferior a 2 y una $RV-$ superior a 0,5 serían de baja utilidad clínica (4).

Ejemplo

Retomando el ejemplo de la tabla 13-2 de un test de antígenos para el diagnóstico de COVID-19, la razón de verosimilitud positiva se calcularía de la siguiente manera:

$$RV+ = \frac{0,56}{1 - 0,99} = 56$$

Este resultado indica que es 56 veces más probable que una prueba sea positiva en un sujeto enfermo que en uno sano.

Y la razón de verosimilitud negativa se calcularía como:

$$RV- = \frac{1 - 0,56}{0,99} = 0,44$$

Este resultado indica que es 2,27 ($1/RV-$) veces más probable un resultado negativo entre los que no tienen la enfermedad que entre los que la tienen.

Las razones de verosimilitud permiten conocer como el resultado de una prueba diagnóstica predice el riesgo de anormalidad mediante la relación de las probabilidades condicionales, es decir, permite relacionar la probabilidad pre-test (antes de la prueba diagnóstica) con la probabilidad post-test (después de la realización de la prueba diagnóstica).

La probabilidad pre-test es la probabilidad de que un sujeto tenga la enfermedad antes de la realización de la prueba diagnóstica, es decir, la prevalencia.

La probabilidad post-test es la probabilidad de que el paciente tenga la enfermedad en un escenario clínico determinado tras la realización de la prueba diagnóstica. La probabilidad post-test depende directamente de la probabilidad pre-test y de las razones de verosimilitud. Para el cálculo de la probabilidad post-test se necesita calcular previamente el Odds pre-test:

$$\text{Odds pre-test} = \frac{\text{prevalencia}}{(1 - \text{prevalencia})}$$

Y el Odds post-test, que se calcula como:

$$\text{Odds post-test} = \text{odss pre-test} \times \text{Razón de verosimilitud}$$

La probabilidad post-test se calcula de la siguiente manera:

$$\text{Probabilidad post-test} = \frac{\text{Odss post-test}}{(1 + \text{Odds post-test})}$$

Ejemplo

Retomando el ejemplo del test de antígenos para el diagnóstico de la COVID-19, que presentaba una RV+ de 56 y una RV- 0,44 y suponiendo una prevalencia de COVID-19 del 9,9% (3).

$$\text{Odds pre-test} = \frac{0,09}{(1 - 0,09)} = 0,10$$

$$\begin{aligned} \text{Odds post-test (positivo)} &= 0,10 \times 56 = 5,6 \\ \text{Odds post-test (negativo)} &= 0,10 \times 0,44 = 0,04 \end{aligned}$$

$$\begin{aligned} \text{Probabilidad positiva post-test} &= 5,6 / (1 + 5,6) = 0,85 \\ \text{Probabilidad negativa post-test} &= 0,04 / (1 + 0,04) = 0,04 \end{aligned}$$

Estos resultados de probabilidad post-test se interpretarían como que la realización de la prueba ha modificado las probabilidades de ser positivo de forma que si antes de realizarse la prueba todos los pacientes tenían una probabilidad de estar infectados de 9,9 %, los que han dado positivo han aumentado su probabilidad hasta el 85 % y los que han dado negativo han disminuido su probabilidad hasta el 4 %.

2.3. El umbral de positividad de una prueba. La curva de características operativas (curva ROC)

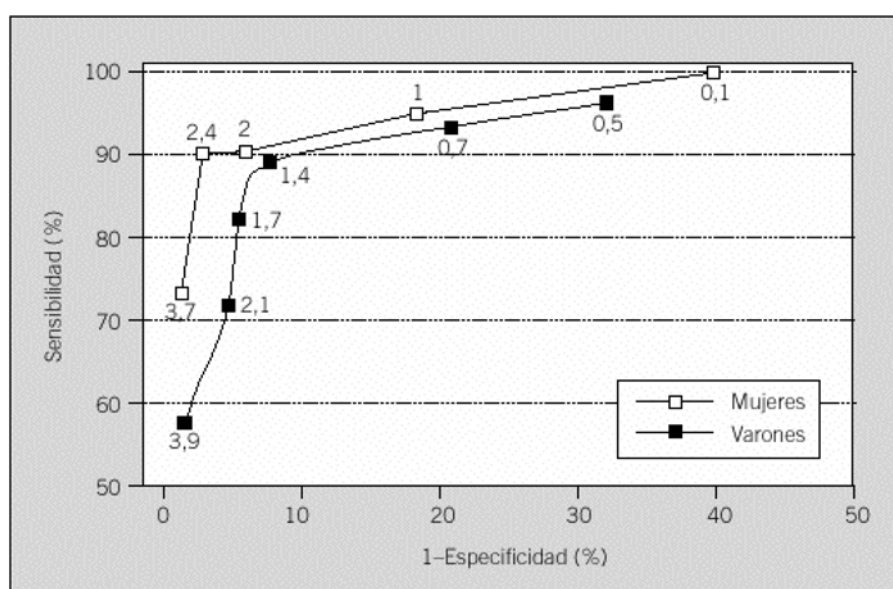
La aplicación de una prueba diagnóstica implica tomar una serie de decisiones que tendrán consecuencias en su evaluación. Un test no aporta la misma información diagnóstica si el umbral para considerar la prueba positiva es bajo que si es muy exigente. La glucemia como prueba diagnóstica de diabetes presenta distintos

valores de sensibilidad y especificidad según el umbral diagnóstico sea 100 o 180 mg/dl. Si se considera este último umbral, casi todas las personas con una prueba positiva (>180 mg/dl) estarán enfermas. Sin embargo, muchas otras personas con diabetes no serán calificadas como enfermas con una prueba tan rigurosa. La prueba sería entonces muy específica pero poco sensible.

Situándose en el otro extremo, y considerando que todos los pacientes con una glucemia mayor de 70 mg/dl fueran diagnosticados como diabéticos, muy pocas personas con la enfermedad dejarían de ser diagnosticadas, pero la mayoría de la gente normal sería etiquetada como diabética. La prueba sería entonces muy sensible pero poco específica. Este balance entre sensibilidad y especificidad debe considerarse a la hora de definir el umbral adecuado para considerar positiva la prueba. La construcción de una curva ROC (Receiver Operator Characteristic Curve) es una herramienta muy útil en esa decisión.

Esta curva se construye con los diferentes resultados que se obtienen en una prueba diagnóstica al colocar el umbral de positividad en diferentes puntos. En la curva se representan en un gráfico los pares sensibilidad y 1-especificidad, que en definitiva son las razones de máxima verosimilitud de un resultado positivo. El eje de abscisas crece conforme disminuye la especificidad y el eje de ordenadas crece conforme aumenta la sensibilidad. En la figura 1 se muestra un ejemplo de curva ROC para una prueba diagnóstica de nefropatía diabética, el cociente albúmina/creatinina en orina.

Figura 1. Curva ROC del cociente albúmina/creatinina.



Fuente: Mundet X, et al. Utilidad del cociente albúmina/creatinina en el diagnóstico de la nefropatía en pacientes con diabetes mellitus tipo 2. *Med Clin (Barc)* 2001; 116: 732-733. Imagen reproducida con autorización del titular de los derechos de explotación de la misma.

El eje de abscisas (1-especificidad) corresponde a la tasa de falsos positivos y el eje de ordenadas (sensibilidad) corresponde a la tasa de verdaderos positivos. Los mejores resultados corresponden a los puntos más cercanos al ángulo superior izquierdo, en el que hay una alta tasa de verdaderos positivos y una baja tasa de falsos positivos. La curva con peor resultado sería una línea diagonal que uniera el extremo inferior izquierdo (valor 0-0) con el superior derecho (valor 100-100), indicando que siempre hay los mismos resultados positivos verdaderos que falsos. En ese caso, la razón de máxima verosimilitud de resultado positivo es 1 en todos los puntos de la curva.

La figura 1 muestra los diferentes pares de datos sensibilidad/1-especificidad del Cociente Albuminuria/Creatinina para cada uno de los puntos de corte establecidos. El punto de corte o umbral de positividad que más se aproxima al extremo superior izquierdo es el que más área deja debajo de la curva, y corresponde con el de mayor razón de máxima verosimilitud. En el caso de las mujeres el mejor punto de corte es 2,4 g/mol, y corresponde con una sensibilidad de 90 % y una especificidad de 97 %. En el caso de los hombres es 1,4 g/mol (sensibilidad 89 % y especificidad 92 %).

Además de poder realizar un análisis cualitativo visualmente, las curvas ROC permiten un análisis cuantitativo preciso para la evaluación de la exactitud de una prueba diagnóstica. El punto que deja mayor área debajo de la curva es el de mayor razón de máxima verosimilitud, y por lo tanto el que ofrece una mayor exactitud.

Visualmente se puede discriminar cuál es el punto que se aproxima más al extremo superior izquierdo, pero para mayor precisión se calcula el área debajo de la curva. Ésta se interpreta como la probabilidad de clasificar correctamente un par de individuos sano y enfermo, seleccionados al azar de la población objeto de estudio, al aplicarles la prueba diagnóstica en cuestión. Es decir, la probabilidad de que el resultado de la prueba resulte más anormal en el sujeto enfermo que en el sano.

El área debajo de la curva es una medida de la exactitud que permite la comparación estadística entre pruebas diagnósticas. Los valores están comprendidos entre 0,5, el peor resultado posible, y 1, el mejor. Un área debajo de la curva de 0,72 significa que un individuo seleccionado aleatoriamente del grupo de enfermos obtendrá el 72 % de las veces un valor mayor en la prueba diagnóstica que un individuo elegido al azar del grupo no enfermo.

3. Fiabilidad o concordancia de una prueba diagnóstica

Los resultados de las mediciones y observaciones clínicas están influidos por multitud de factores como el ambiente en el que se realiza la medición o la persona que realiza la observación. Hay un cierto grado de variabilidad o error intrínseco en cualquier procedimiento de medición, particularmente cuando el componente principal del proceso es la apreciación subjetiva de un observador. La variabilidad raramente puede ser eliminada totalmente, pero el conocimiento de sus causas y su evaluación cuantitativa es necesario para garantizar la calidad de la información que proporciona la prueba.

El grado de variabilidad puede ser evaluado determinando su validez o exactitud, aspecto abordado anteriormente, y además mediante la evaluación de la fiabilidad o concordancia. La **fiabilidad** es la capacidad de una prueba diagnóstica de obtener resultados comparables si se utiliza en condiciones idénticas. Las pruebas diagnósticas deben tener una elevada fiabilidad, de tal forma que haya concordancia o ausencia de desacuerdo en repetidas observaciones o mediciones. La evaluación de la fiabilidad de una prueba diagnóstica se realiza valorando cada una de las fuentes de variabilidad.

Las diferentes fuentes de variación en las observaciones clínicas pueden resumirse como sigue:

- La **variabilidad intraobservador**: el mismo profesional sanitario puede obtener diferentes resultados al repetir una medición en el mismo paciente debido al cambio en el estado físico del profesional, en el grado de atención u otras condiciones ambientales o psicológicas (por ejemplo: presión por parte de los familiares para emitir rápidamente el diagnóstico).
- La **variabilidad interobservador**: las mediciones realizadas por diferentes profesionales suelen ser diferentes, lo cual puede estar motivado por la formación del profesional, la experiencia, la motivación, la agudeza visual o auditiva y otras características personales que pueden afectar a los resultados de las observaciones.
- La **variabilidad biológica** inherente a la característica que se desea medir. Por ejemplo, la tensión arterial está sujeta a cambios en dependencia de la situación en que se encuentre el paciente en el momento de la toma (ingesta previa, estrés, condición física, etc.).
- La **variabilidad del instrumento de medida**. El propio aparato de medición está sometido a un grado de variación, que puede estar motivado por la calidad de la calibración, las unidades en que se mide o la influencia sobre el propio instrumento de factores ambientales como el nivel de ruido o la iluminación.

- La **variabilidad aleatoria**. Es el grado de variación que no puede atribuirse a los factores anteriores. El uso de los valores promedio de una prueba aplicada de forma repetida aumenta la fiabilidad al disminuir el error aleatorio.

La reducción de la variabilidad hasta valores aceptables debe hacerse actuando sobre las diferentes fuentes de variación identificadas. La estandarización del procedimiento de medición puede contribuir a reducir la variabilidad homogeneizando las condiciones de aplicación del procedimiento diagnóstico o entrenando de forma similar a los distintos observadores. Otra forma de reducir la variabilidad es repitiendo la medición en varias ocasiones para tomar la media de las mismas como valor verdadero del sujeto estudiado.

La evaluación de la fiabilidad de una prueba diagnóstica se realiza valorando cada una de las fuentes de variabilidad. Si lo que se pretende es evaluar la variabilidad interobservador se compararán la concordancia entre los resultados obtenidos por los distintos profesionales, manteniendo constantes el resto de los factores que pueden influir en los resultados.

En dependencia del carácter de la variable que se mide, se utilizará un método de evaluación de la variabilidad o concordancia diferente. Cuando se trate de variables nominales o binarias (enfermo/sano, grupo sanguíneo A/AB/B/O) se utilizará el coeficiente Kappa. Si las variables son ordinales se utilizará el coeficiente Kappa ponderado (leve/moderado/grave). Finalmente, cuando las variables sean cuantitativas se utilizará el coeficiente de correlación intraclase.

3.1. El coeficiente de concordancia Kappa

Para desarrollar este concepto se utilizará un ejemplo ilustrativo adaptado de un estudio de evaluación del valor de la tomografía axial computarizada torácica (TACT) para el diagnóstico de COVID-19 (6). Un cálculo sencillo del coeficiente de concordancia Kappa se plantea mediante una tabla de 2x2 en la que se muestran las valoraciones de la TACT por parte de dos observadores (A y B) de forma binaria, es decir positivo (+) o sospechoso de COVID-19, y negativo (0) o no sospechoso. En la tabla 5 se observa una frecuencia muy similar en las valoraciones realizadas por los dos observadores. Sin embargo, no sabemos si los resultados de las observaciones de los clínicos coinciden en los mismos pacientes. Si se describen las lecturas de forma completa mediante una tabla cruzada (tabla 6), se observa que la concordancia no es completa. La proporción de acuerdo se calcula mediante el Índice de Acuerdo Global (IAG), sumando los valores de las celdas en la que los observadores han coincidido, dividido por el número total de observaciones.

Tabla 5. Resultados de las observaciones de dos sanitarios sobre una misma tomografía computarizada torácica para el diagnóstico de COVID-19

	COVID-19		TOTAL
	0	+	
Observador A	48	62	110
Observador B	47	63	110

En la tabla 6 el $IAG = \frac{37+52}{110} = 0,81$. Lo que significa que hay coincidencia entre los dos observadores en el 81 % de los casos.

Tabla 6. Valoración de sospecha de COVID-19 tras la valoración de tomografías computarizadas torácicas para el diagnóstico de COVID-19

		Observador A		
		0	+	Total
Observador B	0	37	10	47
	+	11	52	63
	Total	48	62	110

Sin embargo, no sabemos cuál habría sido el grado de coincidencia por el simple azar. El cálculo del coeficiente de concordancia Kappa sirve para conocer el nivel de acuerdo que se obtendría si descontamos el debido al azar. En la tabla 7 se procede a calcular cuáles serían los resultados si la única causa de la distribución de los resultados obtenidos por los observadores A y B fuese el azar o, lo que es lo mismo, si los sucesos fuesen independientes. El resultado es el Índice de Acuerdo atribuible al Azar (IAA), que resulta un 0,51.

Tabla 7. Frecuencias esperadas por azar en la evaluación de tomografías computarizadas para el diagnóstico de COVID-19^a

		Observador A		
		0	+	Total
Observador B	0	20,51	26,47	47
	+	27,49	35,51	63
	Total	48	62	110

^a Se asume que los sucesos son independientes. En ese caso la probabilidad de que se den dos eventos a la vez es la multiplicación de sus probabilidades. La probabilidad de que se den a la vez los sucesos A0 y B0 es $P(A0) * P(B0)$. Por tanto, la probabilidad de ocurrencia del cuadrante superior izquierdo será $P(A0) * P(B0)$, es decir $(47/110) * (48/110)$, cuyo resultado es un 0,18645, que aplicado a un total de 110 pacientes resulta en 20,51. En el caso del cuadrante inferior derecho podemos hacer un cálculo

simplificado; $P(A+*B+) = \frac{62}{110} * 63 = 35,51$ que se obtiene de la multiplicación de los marginales de la fila y la columna en cuestión, dividido por el número total de observaciones (110). Los dos cuadrantes restantes se calculan restando del total de su fila.

Finalmente el IAA será de 0,51. $IAA = \frac{20,51 + 35,51}{110} = 0,51$

Si al Índice de Acuerdo Global (IAG) le descontamos el Índice de Acuerdo atribuible al Azar (IAA) quedaría sólo el nivel de concordancia que no puede ser atribuida al azar. En nuestro ejemplo sería $0,81 - 0,51 = 0,30$. Este 30 % es el acuerdo observado no debido al azar y constituye el numerador del coeficiente Kappa. Para estimar el denominador se calcula la cantidad máxima de acuerdo que podría alcanzarse no debido al azar, es decir el opuesto del IAA. De este modo, el coeficiente Kappa se expresaría:

$$Kappa = \frac{IAG - IAA}{1 - IAA}$$

En nuestro ejemplo $Kappa = 0,30 / 0,49 = 0,61$. Por tanto, el 61 % de las observaciones son coincidentes una vez eliminado el componente del azar.

Los resultados del coeficiente Kappa varían desde -1, que corresponde con el mayor desacuerdo, a +1, que corresponde con el acuerdo perfecto. El valor 0 correspondería al acuerdo debido al azar.

La fiabilidad se considera pobre si el coeficiente Kappa es menor de 0,2; débil si está entre 0,21 y 0,4, moderada si está entre 0,41 y 0,6, buena si se sitúa entre 0,61 y 0,8 y excelente si es superior a 0,8.

El coeficiente **Kappa Ponderado** se utiliza cuando las variables son ordinales, de tal forma que haya más de dos categorías de valoración con un orden jerárquico entre ellas. Se calcula de forma análoga al coeficiente Kappa Simple, pero multipli-

cando el valor de cada casilla por un peso que pondere la magnitud del desacuerdo, ya que no es lo mismo discrepar entre leve y moderado, que entre leve y grave.

3.2. El Coeficiente de Correlación Intraclass

El Coeficiente de Correlación Intraclass (CCI) se utiliza para valorar la coincidencia cuando las variables medidas son cuantitativas. Es incorrecto utilizar el Coeficiente de Correlación de Pearson (CCP) u otros índices de tendencia para evaluar la fiabilidad.

El siguiente ejemplo ilustra las diferencias entre el CCP y el CCI. En el ejemplo dos observadores toman la tensión arterial sistólica de forma consecutiva y con el mismo esfigmomanómetro a 10 pacientes. Como puede apreciarse en la tabla 8, las mediciones de los dos observadores son diferentes. Los valores registrados por el observador B son en todos los casos superiores a los registrados por el A, y además la diferencia es de igual magnitud casi siempre (en 8 de las 10 ocasiones la diferencia es de 5 mm de Hg, y en las 2 restantes de 10 mm de Hg). Las diferencias son sistemáticas, van siempre en la misma dirección y son de similar magnitud. En este caso el CCP detecta una alta concordancia, concretamente de 0,95. Sin embargo, el CCI sí que detecta que las diferencias son sistemáticas y el resultado refleja una concordancia más real, concretamente de 0,64.

Tabla 8. Medición de la tensión arterial sistólica

Paciente	Observador A (mm Hg)	Observador B (mm Hg)	Diferencia (B-A)
1	135	140	5
2	140	145	5
3	130	135	5
4	145	150	5
5	140	145	5
6	150	160	10
7	140	145	5
8	135	140	5
9	140	145	5
10	135	145	10
Media	139	145	6
Desviación Estándar	5,68	6,67	2,11

El cálculo del CCI se basa en un modelo de análisis de la varianza de medidas repetidas. Se descompone la variación total observada (varianza) en los siguientes componentes: la variabilidad debida a las diferencias entre pacientes o individuos observados (Varianza entre-individuos = σ_E); la varianza intra-individuos (σ_I), en este caso debida a diferencias en mediciones repetidas realizadas por distintos observadores, y por último la varianza residual (σ_R), que es la parte de variabilidad inexplicable y que atribuimos al azar. El CCI se define como el cociente entre la variabilidad entre-individuos y la variabilidad total $CCI = \frac{\sigma_E}{\sigma_E + \sigma_I + \sigma_R}$

Cuando una prueba diagnóstica sea muy fiable o cuando la concordancia entre dos observadores sea muy alta, los resultados de las mediciones repetidas sobre un mismo individuo serán muy parecidos y, en consecuencia, la varianza intra-individuos (σ_I) será muy baja y el valor del CCI será alto. Cuanto más se aproxime a 0 el valor de la varianza intra-individuos, más se acercará a 1 el valor del CCI y viceversa. Puesto que es una proporción, los valores del CCI pueden variar de 0 a 1, donde 0 indica ausencia de concordancia y 1 concordancia o fiabilidad absoluta. Como norma general, valores inferiores a 0,4 indican escasa fiabilidad, entre 0,4 y 0,75 la fiabilidad es buena, y valores superiores a 0,75 reflejan un grado excelente de fiabilidad.

4. Lectura crítica de estudios diagnósticos

En una lectura crítica de un estudio de pruebas diagnósticas tres son las preguntas básicas a contestar, según la propuesta de Evers (7):

- ¿Las evidencias son válidas?
- ¿Discrimina la prueba con precisión entre los pacientes que sí tienen un trastorno específico y los que no lo tienen?
- ¿Se puede aplicar esta prueba diagnóstica a mi contexto?

Los estudios diagnósticos pueden presentar riesgo de sesgos, sobre todo relacionados con deficiencias metodológicas, el reclutamiento de pacientes, la recopilación de datos y la ejecución o interpretación de la prueba.

Los estudios publicados deben contener una descripción completa de los resultados de la investigación y toda la información necesaria para poder juzgarlos. Sin embargo, varios estudios han señalado que las publicaciones sobre estudios diagnósticos a menudo no describen de manera adecuada elementos esenciales, como pueden ser los pacientes incluidos o el diseño de estudio (8).

Para facilitar la realización de un informe completo y transparente de los estudios de exactitud diagnóstica, se desarrolló la declaración STARD (Normas para la presentación de estudios de precisión diagnóstica) inspirada en los Estándares Consolidados para la Comunicación de Ensayos o declaración CONSORT. STARD contiene una lista de verificación de los elementos que deben ser reportados en cualquier estudio de exactitud diagnóstica (8).

La lista de verificación STARD contiene 15 elementos relacionados con la estructura de un estudio científico: título, resumen, introducción, metodología (ej., diseño del estudio, participantes, métodos de análisis), resultados, discusión y otra información, como las fuentes de financiación del estudio (8).

Por otro lado, existen herramientas para analizar la calidad y el riesgo de sesgo de los estudios de exactitud diagnóstica. Entre las herramientas disponibles, destaca la QUADAS 2 (Quality Assessment of Diagnostic Accuracy Studies). Esta herramienta permite analizar el grado de riesgo de sesgo y la aplicabilidad de los resultados del estudio a la pregunta de investigación o contexto analizado (9). QUADAS-2 está dividida en cuatro dominios: los pacientes seleccionados, la prueba índice (prueba diagnóstica analizada), el estándar de oro con el que se compara y el flujo de los pacientes durante el estudio, considerando el tiempo transcurrido entre la realización de la prueba índice y el estándar de oro o referencia.

Un ejemplo de estudio de pruebas diagnósticas ideal sería aquel que incluye todos los pacientes de forma consecutiva o una muestra aleatoria de los pacientes con sospecha de la enfermedad, donde, tanto la prueba índice como la prueba de referencia se realizan al mismo tiempo. Además, la interpretación de la prueba índice debe realizarse sin conocer los resultados de la prueba de referencia. De la misma manera, la interpretación de la prueba de referencia debe realizarse con cegamiento de los resultados de la prueba índice. El conocimiento de los resultados de la prueba de referencia o de la prueba índice pueden sesgar el análisis. Además, el análisis final de los resultados debe realizarse teniendo en cuenta todos los pacientes incluidos al inicio del estudio, ya que las pérdidas de seguimiento pueden alterar su interpretación.

Hay que tener en cuenta que los objetivos de los estudios de evaluación de pruebas diagnósticas pueden ser muy variados y cada uno de ellos requerirá de distintos tipos de diseño de estudio, con sus métodos, limitaciones y riesgo de sesgo específicos. Los estudios clínicos más simples sobre la exactitud diagnóstica de un test pueden adoptar el diseño de estudios transversales, comparando los resultados frente a un test estándar de referencia. Pero cuando se pretende valorar el impacto en el manejo clínico o en el pronóstico del uso de una prueba diagnóstica y se miden resultados relevantes de la condición clínica, los diseños que más información aportan sobre la utilidad clínica de un test diagnóstico son los estudios de cohortes

comparativas y los ensayos aleatorios controlados. Otros aspectos complementarios como el impacto social, ético, organizativo, o como la eficiencia económica de la introducción de un nuevo test diagnóstico pueden exigir el diseño de otro tipo de estudios o la síntesis e integración de información y evidencia procedente de múltiples estudios (revisiones sistemáticas, análisis de decisión clínica, análisis de coste-efectividad).

5. Referencias

1. Knottnerus JA, van Weel C, Muris JWM. Evaluation of diagnostic procedures. *BMJ*. 2002;324(7335):477-80. Available from: <https://pubmed.ncbi.nlm.nih.gov/11859054>
2. Moons KGM, Biesheuvel CJ, Grobbee DE. Test Research versus Diagnostic Research. *Clin Chem*. 2004;50(3):473-6. Available from: <https://doi.org/10.1373/clinchem.2003.024752>
3. Dinnes J, Deeks JJ, Adriano A, Berhane S, Davenport C, Dittrich S, *et al.* Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database Syst Rev*. 2020;8(8):CD013705. Available from: <https://doi.org/10.1002/14651858.CD013705>
4. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ*. 2004;329(7458): 168-9. Available from: <http://www.bmj.com/content/329/7458/168.abstract>
5. Ministerio de Sanidad, Consejo Interterritorial del SNS, Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación. Estudio ENE-COVID: Cuarta ronda del estudio nacional de sero-epidemiología de la infección SARS-CoV-2 en España [Internet]. 2020. Available from: <https://www.mscbs.gob.es/gabinetePrensa/notaPrensa/pdf/15.12151220163348113.pdf>
6. Debray MP, Tarabay H, Males L, Chalhoub N, Mahdjoub E, Pavlovsky T, *et al.* Observer agreement and clinical significance of chest CT reporting in patients suspected of COVID-19. *Eur Radiol*. 2021;31(2):1081-9. Available from: <https://doi.org/10.1007/s00330-020-07126-8>
7. Evers JLH, Land JA, Mol BW. Evidence-based medicine for diagnostic questions. *Semin Reprod Med*. 2003;21(1):9-15.

8. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, *et al.* STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6(11):e012799. Available from: <http://bmjopen.bmj.com/content/6/11/e012799>
9. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-36.

Capítulo 14.

El cribado en la salud pública

ELENA PLANS BERISO Y FERNANDO GARCÍA LÓPEZ

1. Introducción y definición

2. Principios generales del cribado poblacional

3. Características de las pruebas de cribado: validez, fiabilidad y utilidad de las pruebas de cribado

4. Evaluación de la eficacia del cribado

5. Daños potenciales ocasionados por el cribado

6. Aspectos organizativos y económicos

7. Aspectos éticos

8. Conclusiones

Agradecimientos

9. Referencias

Anexo: El cribado en el cáncer: organismos de referencia y ejemplos de balance entre riesgos y beneficios

1. Introducción y definición

El cribado es una intervención de salud pública cuyo objetivo es la detección de una enfermedad no diagnosticada hasta el momento, mediante la realización de pruebas sencillas. Las pruebas de cribado pueden identificar factores de riesgo, predisposición genética, signos tempranos de enfermedad, etc., que permiten clasificar a los sujetos en dos grupos, las personas que tienen un riesgo alto de tener la enfermedad, en cuyo caso seguirán con el proceso diagnóstico, y las que tienen un riesgo bajo, que no se someterán a más pruebas.

El cribado es una estrategia de prevención secundaria y puede ser oportunista o poblacional. El cribado **oportunista** es una actividad no sistemática que se suele realizar dentro de los servicios de salud a petición del interesado o aprovechando la visita del paciente por otros motivos de **atención sanitaria**. Esta estrategia permite identificar posibles enfermedades en etapas tempranas. El **cribado poblacional** es un programa sistemático dirigido por las autoridades sanitarias, que se ofrece a toda la población diana de manera protocolizada y con una evaluación continua de su calidad y sus resultados. En el contexto español, generalmente las **consejerías de Sanidad** de cada comunidad autónoma invitan a las personas objeto del cribado a participar en él, mediante el envío de cartas informativas a sus domicilios, por ejemplo. Este enfoque busca alcanzar a toda la población diana y, al no depender de que la persona visite al médico, se reducen las desigualdades socioeconómicas en salud, con respecto al cribado oportunista.

El cribado oportunista no debe plantearse como una alternativa al poblacional cuando este no ha demostrado ser eficaz. El cribado oportunista solo tendría sentido cuando un cribado que se ha mostrado eficaz a nivel poblacional no está siendo aplicado por razones económicas, logísticas, organizativas o de cualquier otro orden.

Como sinónimos de cribado se usan los términos *detección precoz* o *despistaje*.

2. Principios generales del cribado poblacional

2.1. Criterios de la enfermedad

Para entender este apartado necesitamos entender el concepto de *historia natural de la enfermedad*. En las figuras 1 y 2, se muestran dos ilustraciones que representan la historia de la enfermedad con el ejemplo de cribado en el caso del cáncer (figura 1) y del cribado neonatal (figura 2).

La historia natural de la enfermedad describe cómo una enfermedad se desarrolla y progresa en los individuos de una población. Este proceso consta de diferentes etapas: en primer lugar, algunas personas son más susceptibles a la enfermedad que otras y, después, puede iniciarse la enfermedad sin manifestar síntomas, lo que se llama *fase preclínica*. Durante esta fase preclínica, hay un punto a partir del cual, si se realizara una prueba de cribado, la enfermedad sería detectable, lo que define un periodo llamado *fase preclínica detectable*. Por otro lado, el *punto crítico* en la historia natural de la enfermedad se refiere al momento de *no retorno* en el cual ocurre un cambio significativo en la evolución de la enfermedad. Es el punto de inflexión en el que la enfermedad progresa de una etapa a otra, como el paso de una fase asintomática a una fase sintomática o el inicio de complicaciones graves. El punto crítico puede ser determinante en el pronóstico de la enfermedad y en las opciones de tratamiento disponibles. Cuando se realiza un cribado, se busca detectar la enfermedad en el periodo de ventana que se encuentra entre el inicio de la fase preclínica detectable y el punto crítico.

Figura 1. Cribado en el cáncer.

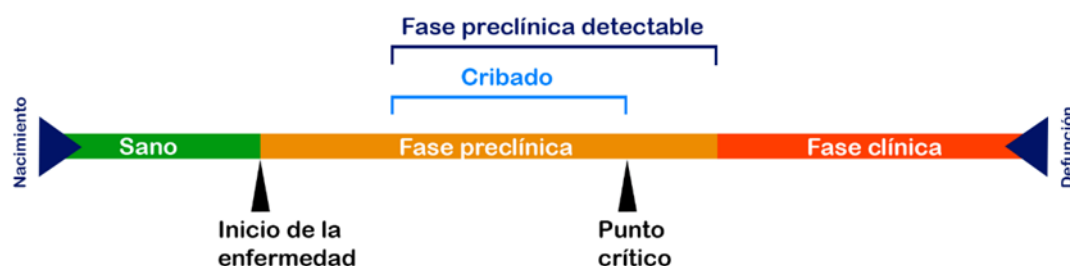


Figura 2. Cribado neonatal.



La enfermedad que se va a cribar debe constituir un problema de salud importante, es decir, que afecte a un grupo amplio de personas y que produzca una morbilidad o una mortalidad considerables. Por otra parte, debe tener una his-

toria natural bien conocida, lo cual implica que se comprendan los factores que contribuyen a su desarrollo y progresión. Además, debe tener una fase preclínica detectable suficientemente prolongada, en la cual no se presenten síntomas graves o típicos de la enfermedad, pero se puedan detectar factores de riesgo, precursores, biomarcadores, síntomas leves o inespecíficos u otros signos tempranos de la enfermedad.

2.2. Criterios de la prueba de cribado

La prueba debe estar validada, lo que significa que ha sido sometida a rigurosos estudios y análisis para demostrar su validez y fiabilidad.

Es fundamental conocer bien la distribución de los resultados de la prueba en la población diana, tanto en personas sanas como en aquellas con la enfermedad que se busca diagnosticar.

Además, en las pruebas basadas en variables continuas, debe estar bien definido y acordado el punto de corte o umbral a partir del cual se considera que una persona tiene un resultado positivo, indicador de la posible presencia del proceso o de la enfermedad en cuestión. Es importante que este punto de corte esté respaldado por suficiente evidencia científica y sea ampliamente aceptado por la comunidad médica. Como se expone en el apartado 3.1. de este capítulo, este punto de corte ha de tener en cuenta el contexto de la prueba de cribado y puede cambiar para ser, por ejemplo, más sensible.

La prueba debe ser sencilla y fácil de entender y de aplicar tanto para los profesionales de la salud como para los pacientes.

La prueba debe ser segura tanto para quienes la realizan como para quienes se someten a ella. Esto implica tener en cuenta cualquier posible riesgo o efecto adverso asociado y garantizar que se toman las precauciones necesarias para minimizarlos.

Por último, es necesario tener una política acordada sobre las pruebas diagnósticas que se deben realizar en los casos con resultados positivos para confirmar el diagnóstico, evaluar la extensión de la enfermedad y elegir las opciones terapéuticas. Estos algoritmos de decisión clínica deben basarse en la evidencia científica disponible y en las guías de práctica clínica, siempre considerando aspectos como los efectos secundarios y las preferencias del paciente.

2.3. Criterios del tratamiento de la enfermedad

Para recomendar una prueba de cribado es necesario que exista un tratamiento eficaz disponible, que disminuya la morbilidad resultante de la enfermedad que se quiere cribar en comparación con el curso natural de la enfermedad sin cribado.

Asimismo, es necesario establecer expectativas razonables. Esto implica que los individuos deben estar informados sobre la importancia de seguir las recomendaciones del programa de cribado, incluido el seguimiento y tratamiento adecuado de las lesiones encontradas.

2.4. Criterios del programa de cribado

El cribado poblacional debe respaldarse con pruebas sólidas que demuestren su efectividad en la reducción de la mortalidad o la morbilidad. En el caso en que el cribado solo proporcione información para facilitar una *elección informada*, como en el síndrome de Down o en portadores de fibrosis quística, es crucial contar con suficientes pruebas científicas de que el programa de cribado mide el riesgo asociado con validez y fiabilidad.

Para respaldar la implementación de un programa de cribado, se requieren pruebas provenientes de revisiones sistemáticas con metanálisis de ensayos controlados y aleatorizados de alta calidad, que demuestren de manera concluyente la reducción de la mortalidad o la morbilidad y a quiénes se les debe ofrecer este programa.

Es fundamental contar con suficientes pruebas que demuestren que el programa de cribado completo es aceptable clínica, social y éticamente tanto para los profesionales de la salud como para la población en general. Esto asegura que los profesionales estén dispuestos a implementar el programa y que la población diana esté dispuesta a participar activamente en el cribado.

Además, los beneficios obtenidos por los individuos a través del cribado deben superar claramente los inconvenientes del programa. Estos inconvenientes, que serán tratados más adelante, pueden incluir el sobrediagnóstico, los hallazgos incidentales (*incidentalomas*), el sobretratamiento, los falsos positivos, los errores de clasificación, la falsa seguridad y las complicaciones. Es esencial que la relación entre beneficios e inconvenientes sea cuidadosamente evaluada y que los beneficios obtenidos justifiquen los riesgos.

Asimismo, el coste de oportunidad del programa de cribado debe estar equilibrado económicamente en relación con los gastos totales de la asistencia sanitaria. Este coste de oportunidad es lo que se podría haber hecho con ese mismo presu-

puesto si no se hubiera invertido en el cribado. Es necesario evaluar y considerar los recursos disponibles y asegurarse de que el programa sea económicamente viable y sostenible y se eviten desequilibrios financieros en el sistema de atención sanitaria.

2.5. Principios de Wilson y Jungner

Los criterios expuestos desde el apartado 2.1. hasta el 2.4. se alinean con los principios de Wilson y Jungner (Tabla 1).

Wilson y Jungner fueron dos médicos suecos que realizaron importantes aportaciones en el campo de la salud pública y la epidemiología. Sus contribuciones han sido fundamentales para garantizar que los programas de cribado sean efectivos, éticamente responsables y beneficiosos para la población.

Tabla 1. Principios de Wilson y Jungner

Principios de Wilson y Jungner	Descripción
Importancia del problema de salud	El problema de salud debe ser relevante y tener un impacto significativo en la población.
Criterios de la enfermedad	Los criterios para definir la enfermedad deben estar claramente establecidos y ser comprensibles.
Naturaleza de la enfermedad	Debe conocerse la historia natural de la enfermedad, incluyendo su etapa preclínica y sintomática.
Disponibilidad de un tratamiento	Deben existir tratamientos efectivos y aceptables para la enfermedad en la etapa temprana.
Detección en la etapa preclínica	Debe ser posible detectar la enfermedad en su fase preclínica, antes de que aparezcan síntomas.
Prueba de cribado aceptable	El cribado debe ser una prueba confiable, segura, accesible y aceptable para la población.
Programa de diagnóstico temprano	Debe existir un protocolo establecido para el diagnóstico temprano y el seguimiento de los casos.
Aceptabilidad para la población	El cribado debe ser aceptado por la población objetivo y por los profesionales de la salud.
Intervención coste-efectiva	El cribado y el tratamiento temprano deben ser coste-efectivos para el sistema de salud.
Evaluación continua del programa	Es necesario evaluar periódicamente la eficacia y eficiencia del programa de cribado.
Consideraciones éticas	Se deben tener en cuenta los aspectos éticos, incluida la privacidad y el consentimiento informado.

3. Características de las pruebas de cribado: validez, fiabilidad y utilidad de las pruebas de cribado

En el capítulo de estudios diagnósticos se explican la validez, la fiabilidad y los valores predictivos de las pruebas diagnósticas. Como recordatorio breve, es importante tener en cuenta que, aunque la sensibilidad y la especificidad son los parámetros que definen la exactitud de una prueba, los valores predictivos positivo y negativo son los que permiten evaluar el rendimiento de la prueba diagnóstica en una población específica con una cierta proporción de personas enfermas.

No obstante, es importante resaltar que las pruebas de cribado tienen sus particularidades. En general, las pruebas de cribado deben dar prioridad a la sensibilidad respecto a la especificidad ya que es importante no dejar posibles casos fuera. Sin embargo, esta premisa está sujeta al contexto económico y social y a las características de la enfermedad, entre otras consideraciones. Tomemos como ejemplo el sida, una enfermedad estigmatizada en determinados contextos sociales: resulta crucial que la prueba de cribado cuente no solo con una alta sensibilidad, sino con una elevada especificidad para prevenir falsos positivos y así evitar que las personas sufran el estigma asociado a esa enfermedad de manera injustificada.

En cuanto a la utilidad de la prueba de cribado, en líneas generales, es importante que tenga un **valor predictivo positivo alto**. El valor predictivo de una prueba positiva es la probabilidad de que una persona tenga la enfermedad si ha tenido un resultado positivo.

Cuando se aplica a una población una prueba de cribado con un valor predictivo positivo alto, se maximiza la eficacia clínica y la eficiencia económica al concentrar los recursos y los esfuerzos en las personas que tienen más probabilidades de beneficiarse de la detección temprana. Esto a su vez reduce los costes asociados con la práctica de pruebas adicionales, tratamientos innecesarios y el hallazgo de falsos positivos. Por tanto, es importante dirigir las pruebas de cribado hacia las poblaciones de mayor riesgo, teniendo en cuenta factores como la edad, los antecedentes familiares y otros factores de riesgo conocidos, ya que el valor predictivo positivo aumenta con la prevalencia de la enfermedad en la población diana.

4. Evaluación de la eficacia del cribado

La evaluación de la eficacia del cribado es un proceso crucial para determinar su impacto y utilidad. Una forma de evaluar la eficacia del cribado es a través de

medidas de desenlace y de medidas operativas. Estas medidas proporcionan una evaluación cuantitativa y objetiva de los resultados obtenidos mediante el cribado, tanto en términos de los beneficios para la salud como de la eficiencia operativa del programa.

4.1. Medidas de desenlace

Estas medidas nos permiten examinar los resultados tangibles y los impactos en la salud de la población sometida al cribado. A continuación, se presentan algunas medidas de desenlace clave en la evaluación del cribado.

1. **Reducción de la mortalidad.** La meta principal del cribado es reducir la mortalidad por cualquier causa. La medida de desenlace es la comparación de las tasas de mortalidad globales entre aquellos sometidos al cribado y aquellos que no lo han sido. Para detectar una posible reducción de la mortalidad general, por todas las causas, se requieren muestras muy grandes, no siempre factibles de obtener en los ensayos controlados y aleatorizados, que utilizan con frecuencia como medida principal de efectividad la mortalidad específica por la enfermedad sujeta a cribado.
2. **Reducción de la letalidad.** Además de la mortalidad general, la reducción de la letalidad es otro indicador importante. La letalidad es la probabilidad de muerte entre los enfermos. Al evaluar el cribado, se busca determinar si hay una disminución en la letalidad en aquellos casos detectados a través del programa.
3. **Aumento del porcentaje de casos detectados en etapas precoces respecto al total de diagnósticos de la enfermedad.** El cribado tiene como objetivo identificar la enfermedad en sus etapas más tempranas, cuando el tratamiento y la intervención pueden tener un mayor impacto.
4. **Reducción de complicaciones.** El cribado exitoso debería llevar a una disminución en las complicaciones asociadas a la enfermedad en los individuos identificados a través del cribado y tratados tempranamente en comparación con los casos de enfermedad que no fueron sometidos al cribado.
5. **Reducción de recidivas o metástasis.** En algunas enfermedades, como el cáncer, el cribado exitoso puede contribuir a la reducción de recidivas (reaparición de la enfermedad) y metástasis (diseminación de la enfermedad a otras partes del cuerpo). Estas medidas son indicadores importantes de la eficacia del cribado para evitar la progresión y la propagación de la enfermedad.

6. **Mejora de la calidad de vida relacionada con la salud de la población.** Además de los resultados clínicos, es esencial evaluar la mejora de la calidad de vida relacionada con la salud de la población sometida al cribado. Esto implica examinar el impacto en la salud percibida de los participantes en el cribado, la reducción de los efectos negativos de la enfermedad y los perjuicios del cribado *per se* (sobrediagnóstico, sobretratamiento, falsos positivos, etc.).

4.2. Medidas operativas

Estas medidas nos brindan información sobre la implementación y el funcionamiento del programa de cribado. A continuación, se presentan algunas medidas operativas clave utilizadas en la evaluación del cribado.

1. **Número de personas sometidas a cribado.** Proporciona una visión general de la cobertura y participación en el programa.
2. **Proporción de la población diana sometida a cribado.** Es importante para evaluar la aceptación y la participación (cobertura) de la población en el programa.
3. **Prevalencia detectada de enfermedad preclínica.** Se refiere a la proporción de individuos sometidos al cribado que se detectan con la enfermedad en su fase preclínica. Esta medida es esencial para evaluar la capacidad del programa de cribado para detectar tempranamente la enfermedad en la población.
4. **Costes totales del programa.** Incluye los costes de las pruebas, el seguimiento, el diagnóstico y el tratamiento de los casos identificados. Evaluar los costes totales ayuda a determinar la eficiencia del programa.
5. **Costes por caso detectado.** Ayuda a evaluar la relación entre los recursos invertidos y el número de casos detectados. Un menor coste por caso detectado indica una mayor eficiencia del programa.
6. **Proporción de personas con cribado positivo en las que se establece diagnóstico definitivo y tratamiento.** Es importante para evaluar la efectividad del programa en la gestión y seguimiento de los casos identificados.
7. **Valores predictivos.** Tanto el valor predictivo positivo como el valor predictivo negativo son medidas que indican la probabilidad de que un resultado positivo o negativo en el cribado sea verdadero.

4.3. Sesgos en la evaluación de la eficacia del cribado

Los estudios que tratan de evaluar la eficacia de un programa de cribado pueden incurrir en una serie de sesgos, que se detallan a continuación.

4.3.1. Sesgos de selección

Son errores sistemáticos derivados de que los sujetos estudiados o participantes no son representativos de la población objetivo, bien por muestreo no aleatorio, en estudios de prevalencia, o bien porque los sujetos de dicha población tienen distinta probabilidad de participar según estén en una u otra categoría de exposición o desenlace, en estudios analíticos (ver capítulo de sesgos).. Es importante tener en cuenta estos sesgos de selección al interpretar los resultados de la evaluación de la efectividad de las pruebas de cribado.

Un sesgo de selección común en la evaluación de la eficacia del cribado es el sesgo de voluntariedad.

El sesgo de voluntariedad ocurre cuando los individuos se autoseleccionan para participar en la prueba de cribado. Esto puede alterar la representatividad, ya que estas personas pueden tener características diferentes a la población general. Por ejemplo, aquellos que optan por someterse al cribado pueden tener un estado de salud distinto a los que no se someten, ya sea porque estén más sanos o porque tengan más riesgo de desarrollar la enfermedad. Esto puede llevar a una detección de enfermedades distinta que en la población objetivo y, por tanto, a resultados clínicos falsamente distintos en comparación con aquellos que no se someten al cribado.

Con frecuencia, los sujetos que se someten al cribado tienen un mayor riesgo de desarrollar su enfermedad objetivo. Por ejemplo, en el caso del cáncer de mama, las mujeres que optan por someterse al cribado pueden tener un riesgo genético o una historia familiar de la enfermedad, lo que puede influir en los resultados clínicos observados.

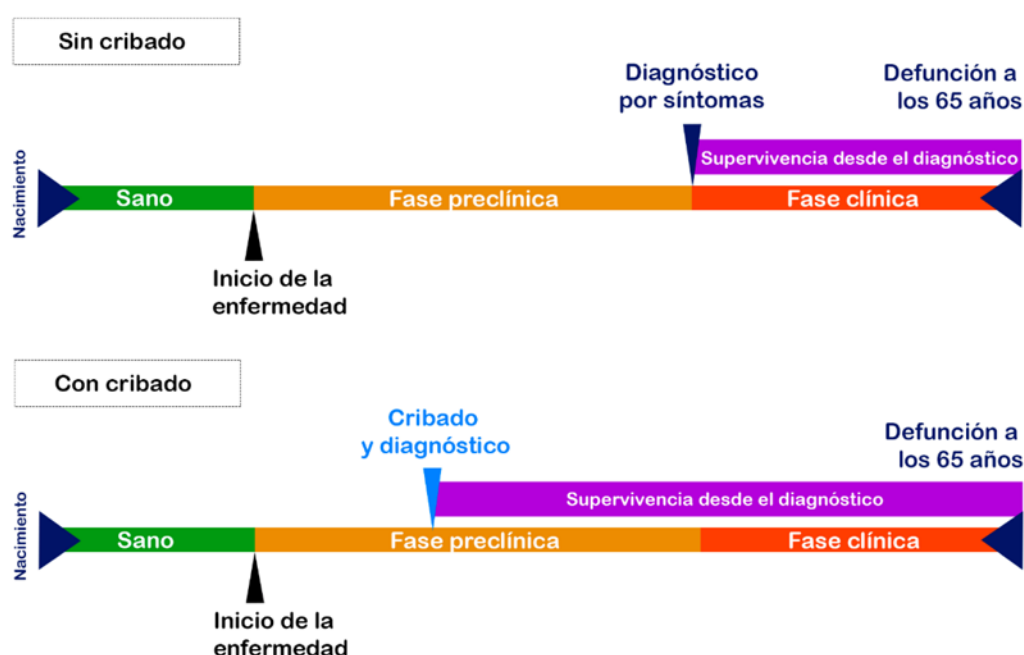
4.3.2. Sesgo de adelanto diagnóstico (*lead time bias*)

Se trata de un sesgo de información que consiste en interpretar erróneamente que un diagnóstico y tratamiento precoz aumentan el tiempo de supervivencia cuando lo que ocurre en realidad es que la presencia de la enfermedad se conoce durante una proporción mayor en la vida del paciente, pero la muerte se produce en el mismo momento, con o sin cribado.

Este sesgo se conoce también con la denominación *periodo de ventaja*, de ventaja por hacerse el diagnóstico a tiempo o *desplazamiento del momento cero*.

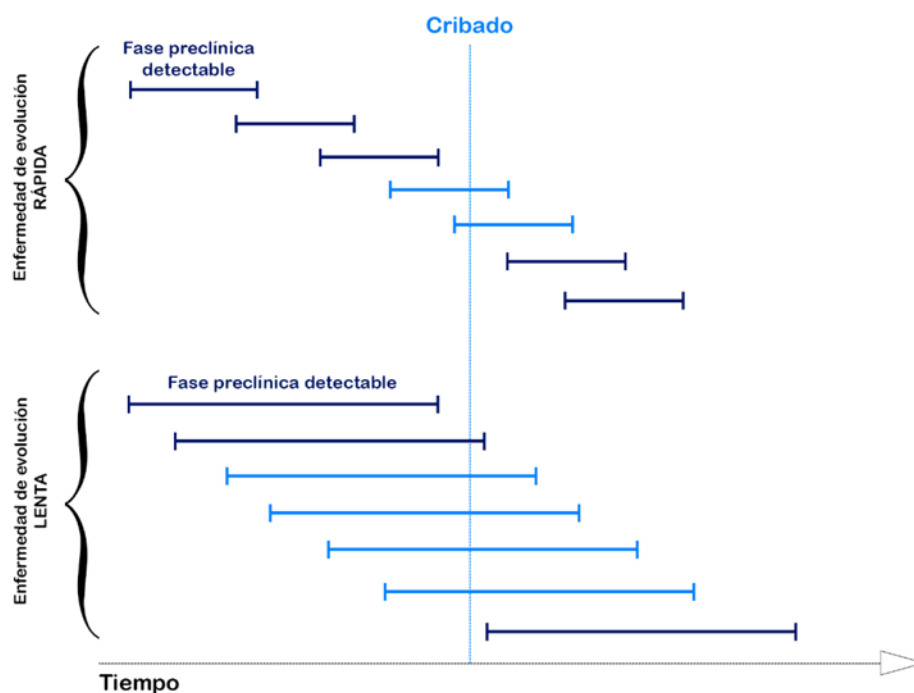
En la Figura 3, se muestran dos casos de la historia natural de una misma enfermedad, uno sin recibir cribado y, por lo tanto, diagnosticado con el inicio de los síntomas, y otro con cribado, diagnosticado antes de que aparezcan los síntomas. A pesar de la diferencia en el momento del diagnóstico, ambos casos presentan la misma edad de defunción (65 años) cuya causa es la misma enfermedad. Esto puede llevar a interpretaciones equivocadas, como pensar que el cribado es beneficioso porque aumenta la *supervivencia desde el diagnóstico*, cuando en realidad lo que ocurre es que el cribado permite detectar la enfermedad antes, mientras que la supervivencia real (no la *supervivencia desde el diagnóstico*) se mantiene igual.

Figura 3. Sesgo del adelanto diagnóstico.



4.3.3. Sesgo de duración (*length bias*)

El sesgo de cribado por duración (*length bias*) es un tipo de sesgo de selección que surge debido a que el cribado suele detectar más fácilmente los casos de la enfermedad que tienen una duración más larga o una progresión más lenta en comparación con aquellos casos que tienen una duración más corta o una progresión más rápida. Esto puede llevar a pensar que el cribado aumenta la supervivencia cuando en realidad lo que hace es encontrar más casos de progresión lenta. No se debe confundir con el tiempo de adelanto diagnóstico (figura 4).

Figura 4. Sesgo de duración.

4.3.4. Sesgo de sobrediagnóstico

Se trata de un sesgo de información, pues clasifica erróneamente casos como positivos cuando su historia natural real será benigna. Este sobrediagnóstico se produce cuando el cribado identifica lesiones o anomalías que nunca se habrían diagnosticado en ausencia del cribado. Es un caso extremo del sesgo de duración y puede llevar a un tratamiento innecesario de afecciones que no habrían causado daño durante la vida del individuo.

Confirmar definitivamente la existencia de este sesgo puede ser difícil, pero se han observado ejemplos que sugieren su presencia. Por ejemplo, el cribado de cáncer de próstata, con el antígeno prostático específico (PSA, por sus siglas en inglés) ha llevado a detectar un porcentaje considerable de casos que nunca habrían presentado síntomas clínicos ni requerido tratamiento. Lo mismo ocurre con casos de carcinoma ductal *in situ* detectados por mamografía y biopsia, así como casos de carcinoma papilar de tiroides detectados por ecografía y biopsia por aspiración con aguja fina.

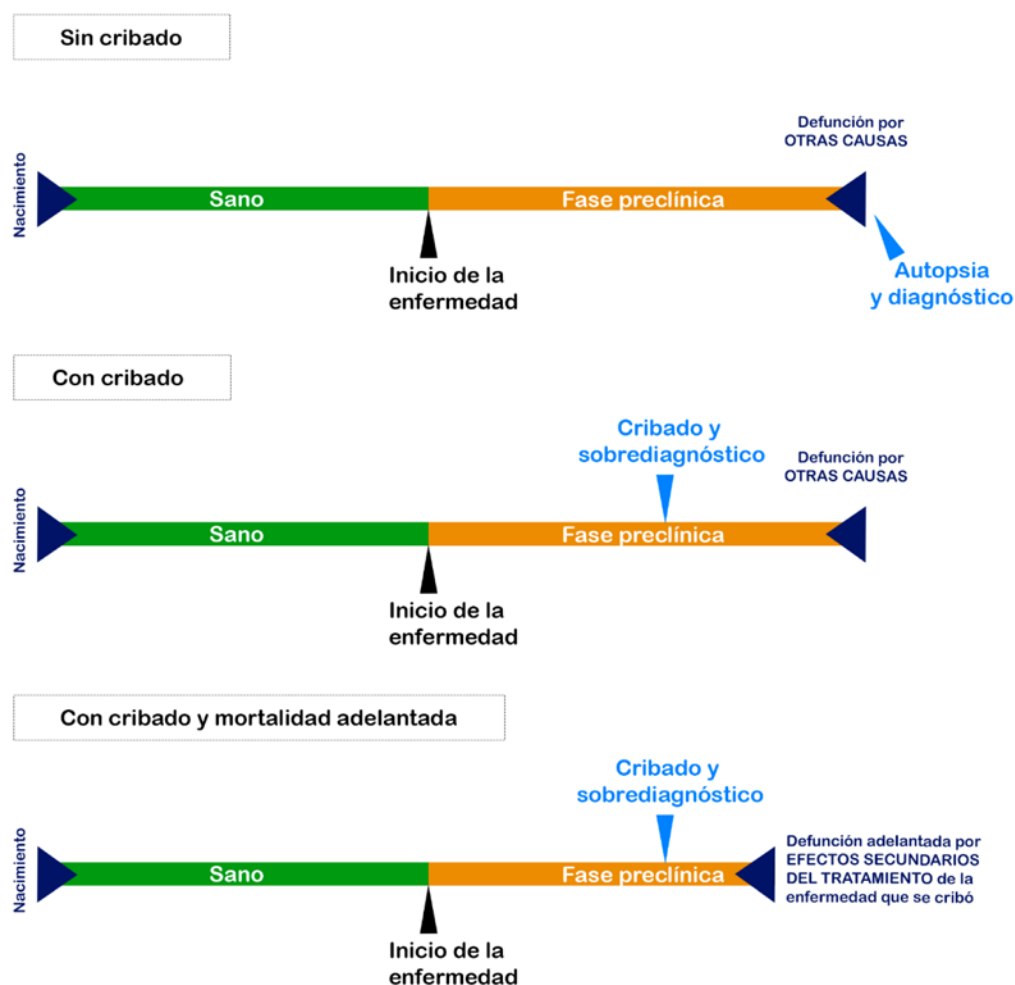
El sobrediagnóstico plantea desafíos importantes en la evaluación del cribado. Puede conducir a tratamientos innecesarios, con sus correspondientes riesgos y efectos secundarios, así como a un aumento en los costes de la atención médica. Además, puede generar ansiedad y angustia en personas que reciben un diagnóstico y tratamiento, que muchas veces tiene complicaciones de salud serias y deterioro de la calidad de vida por una afección que no habría causado daño alguno.

Por definición, la presencia de sobrediagnóstico sobrestima la supervivencia específica del cribado, ya que en los estudios que comparan la supervivencia de las personas sometidas al cribado con la de las que no, el grupo sometido al cribado se compone de casos que tendrán una supervivencia mayor por no desarrollar la enfermedad clínicamente, casos que no llegan a detectarse en el grupo sin cribado.

El sobrediagnóstico plantea un problema importante para los estudios de evaluación de la eficacia del cribado, pues no se puede controlar mejorando el diseño del estudio. Solo se puede controlar mejorando el valor pronóstico de la prueba de cribado. Por ejemplo, la PSA para el diagnóstico del cáncer de próstata no distingue entre los casos progresivos y los de crecimiento lento o sin crecimiento, pero una prueba futura mejor sí podría hacerlo.

Antes de implantar un programa de cribado, es fundamental considerar cuidadosamente sus beneficios y sus riesgos. Se debe evaluar de manera crítica la evidencia científica disponible, sopesar los posibles beneficios de la detección temprana contra los riesgos de sobrediagnóstico y tratamientos innecesarios, y tener en cuenta las preferencias y los valores de los pacientes (figura 5).

Figura 5. Sesgo de sobrediagnóstico.



4.4. Sesgos en la evaluación de la efectividad del cribado

En algunos estudios de la efectividad del cribado, no de su eficacia, se pueden encontrar otros sesgos de selección que se enumeran a continuación:

1. **Sesgo de adherencia.** Se refiere a la mayor tendencia de los participantes que se someten a pruebas de cribado a cumplir con las recomendaciones y tratamientos en comparación con los que no se someten a cribado. Esto puede afectar la estimación de los beneficios y riesgos reales de la prueba.
2. **Sesgo de selección del proveedor de atención sanitaria.** Puede ocurrir cuando los proveedores de atención sanitaria seleccionan a ciertos pacientes para someterse a pruebas de cribado en función de criterios subjetivos o personales.
3. **Sesgo de exclusión.** Se produce cuando ciertos grupos de la población son excluidos de las pruebas de cribado por barreras económicas, culturales o de acceso a los servicios sanitarios, etc. Esto puede limitar la generalización de los resultados si estos son diferentes de los observados la población que recibe el cribado.
4. **Sesgo de referencia.** Ocurre cuando los participantes se seleccionan en función de su estado de salud o de la presencia de síntomas, lo que puede sobrestimar o subestimar los beneficios de la prueba de cribado.

4.5. Otros errores comunes en la valoración de la eficacia de las pruebas de cribado

La interpretación de la eficacia de las pruebas de cribado presenta errores comunes, que incluyen:

- Equiparar la sensibilidad de la prueba de cribado en pacientes en fase preclínica con la misma prueba en pacientes en la fase clínica puede llevar a conclusiones erróneas ya que la sensibilidad de una prueba puede variar según la etapa de la enfermedad en la que se encuentre el individuo. Este error se llama *sesgo del espectro* (Ransohoff y Feinstein, 1978). Por lo tanto, es fundamental evaluar la sensibilidad utilizando una muestra de individuos en la fase preclínica o asintomática, ya que puede diferir de la sensibilidad observada en pacientes sintomáticos.
- Utilizar como patrón de referencia (*gold standard*) un marcador intermedio en vez de un resultado clínicamente establecido de referencia, fiable y reconocido, para evaluar la sensibilidad y especificidad de una prueba de cribado. Por ejemplo, en la evaluación de la sensibilidad y especificidad de la prueba de cribado de cáncer de pulmón con TAC (sin contraste) sería

preferible usar como patrón de referencia la biopsia, y no otra prueba de imagen como podría ser un angio-TAC.

- Si evaluamos la efectividad de una prueba de cribado y hacemos un seguimiento incompleto de los casos, es decir, seguimos a aquellos con resultados positivos en la prueba de cribado, pero dejamos de seguir a los que obtienen resultados negativos antes de confirmar con el patrón de oro si son verdaderos o falsos negativos, es posible que sobrestimemos la sensibilidad de la prueba de cribado. Es fundamental realizar un seguimiento adecuado y comparar todos los resultados negativos con el patrón de oro para obtener una estimación precisa de la sensibilidad de la prueba.

4.6. Diseños de estudio para la evaluación de la eficacia del cribado

Los **ensayos controlados y aleatorizados** constituyen el diseño más sólido para establecer la eficacia del cribado. En primer lugar, los ensayos controlan la confusión, pues la distribución aleatoria permite que el grupo asignado al cribado y el grupo control sean intercambiables, por lo que se podrá establecer una relación causal entre el cribado y los desenlaces que se estudien. Por otra parte, los ensayos controlados y aleatorizados permiten prevenir los sesgos de selección que se han enumerado y el sesgo de adelanto diagnóstico. Sin embargo, para el control del sesgo de duración, el diseño del estudio debe tener en cuenta el ritmo de progresión de la enfermedad, dado que en casos de progresión rápida de la enfermedad el intervalo de los cribados ha de ser menor para detectar todos los casos y no subestimar la capacidad de detección que tiene la prueba de cribado. Por último, en los ensayos controlados y aleatorizados, que no pueden controlar el sobrediagnóstico, deberían cuantificarlo para tenerlo en cuenta en el análisis riesgo-beneficio del programa de cribado.

Aunque los ensayos controlados y aleatorizados son considerados el patrón de referencia en la evaluación de la eficacia del cribado, en determinadas situaciones se ha recurrido a otros diseños de estudio, como los estudios de casos y controles, de cohortes, diseños cuasiexperimentales y ecológicos; además, resultan de utilidad los estudios de compilación, como las revisiones sistemáticas y metanálisis.

5. Daños potenciales ocasionados por el cribado

El cribado, aunque puede ser una herramienta valiosa para la detección temprana de enfermedades, también conlleva una serie de posibles daños y efectos ad-

versos que hay que considerar. A continuación, se detallan algunos de los posibles daños potenciales ocasionados por el cribado:

- **Falsos negativos.** Los falsos negativos pueden llevar a una engañosa sensación de seguridad y retrasar el diagnóstico y tratamiento adecuados.
- **Falsos positivos.** Pueden generar ansiedad y estrés innecesarios, así como pruebas y procedimientos adicionales, que pueden ser invasivos o tener riesgos asociados. Además, los falsos positivos pueden llevar a tratamientos innecesarios y costosos, con efectos secundarios potenciales.
- **Sobrediagnóstico.** Puede llevar a tratamientos innecesarios, con sus correspondientes riesgos y efectos secundarios, así como a un aumento en los costes de la atención médica.
- **Hallazgos incidentales.** Durante el cribado, es posible que se descubran anomalías o afecciones no relacionadas directamente con la enfermedad objetivo del cribado. Estos hallazgos, también llamados *incidentalomas*, pueden requerir pruebas y seguimientos adicionales, y pueden generar ansiedad y preocupación en los individuos y añadir costes y riesgos innecesarios. Un ejemplo de *incidentaloma* sería el hallazgo de una masa en el hígado que se encuentran de manera fortuita durante una ecografía o una tomografía computarizada abdominal y que no corresponde con ningún cáncer o enfermedad de mal pronóstico (hemangioma hepático, adenoma hepático, quiste hepático, etc.), y que conducirá a pruebas adicionales como ecografías, tomografías computarizadas, resonancias magnéticas o incluso biopsias para determinar el diagnóstico.
- **Daños relacionados con el procedimiento de cribado.** Algunos procedimientos derivados del cribado, como biopsias o pruebas invasivas, conllevan riesgos y posibles complicaciones, como infecciones, sangrado o malestar físico.
- **Afectación psicológica.** El proceso de cribado puede tener un impacto emocional en las personas, generando ansiedad, estrés y preocupación ante la posibilidad de recibir un diagnóstico positivo. Incluso en el caso de resultados negativos, algunas personas pueden experimentar un miedo persistente a futuras pruebas o desarrollar una hipocondría relacionada con la enfermedad objetivo del cribado.

Es clave evaluar el beneficio-riesgo del cribado, teniendo en cuenta estos posibles daños y efectos adversos. Se requiere una comunicación clara y una toma de decisiones compartida entre los profesionales de la salud y los individuos sometidos al cribado, considerando tanto los aspectos clínicos como los impactos psicológicos y sociales.

6. Aspectos organizativos y económicos

Es necesario establecer una definición clara de la población objetivo y asegurar la identificación de los sujetos elegibles, así como implementar medidas para garantizar una alta cobertura del programa. También se debe desarrollar un plan integral para gestionar y monitorizar el programa, estableciendo estándares e indicadores de control de calidad.

Se debe contar con personal adecuado para llevar a cabo el cribado, derivar a los participantes, realizar diagnósticos, proporcionar tratamiento y gestionar el programa en su totalidad. Antes del inicio del programa, se debe brindar a los participantes información suficiente sobre los objetivos y posibles beneficios y consecuencias adversas del cribado, que les permita tomar decisiones informadas.

Además, es importante prever las presiones del público para ampliar los criterios de elegibilidad, reducir el intervalo de cribado o aumentar la sensibilidad de la prueba. Las decisiones relacionadas con estos aspectos deben estar debidamente justificadas y comunicadas a la sociedad.

Por último, se debe realizar una evaluación económica exhaustiva del programa para comprender su viabilidad financiera y los beneficios asociados.

7. Aspectos éticos

Los aspectos éticos desempeñan un papel fundamental en el diseño, implementación y evaluación de un programa de cribado. A continuación, se presentan algunos aspectos éticos relevantes que hay que considerar:

1. **El beneficio y el riesgo para la comunidad.** Los beneficios comunitarios esperados del cribado deben superar los riesgos potenciales para justificar su implementación. Se debe tener en cuenta la equidad en la distribución de beneficios y riesgos, así como la necesidad de minimizar cualquier daño o impacto negativo en los individuos sometidos al cribado.
2. **Los riesgos de la prueba de cribado.** Los promotores del programa de cribado tienen la responsabilidad de demostrar la eficacia de los procedimientos y de garantizar que los riesgos y daños asociados sean menores que los de la atención sanitaria estándar para los pacientes de esa enfermedad.

3. **El coste de oportunidad.** Es importante recordar que los recursos son limitados y que un programa de cribado implica el uso de recursos que podrían asignarse a otras opciones preventivas o de asistencia sanitaria. Por tanto, se debe realizar una evaluación cuidadosa de los costes y beneficios del cribado, considerando si es la mejor asignación de recursos en comparación con otras intervenciones o programas que podrían generar mayores beneficios para la salud pública.
4. **El consentimiento informado.** El consentimiento informado, que parte del principio del respeto hacia las personas, es un elemento esencial en el cribado. Los individuos deben recibir una información completa y comprensible sobre la prueba de cribado, las posibles consecuencias del resultado, el eventual proceso diagnóstico y los tratamientos potenciales. Se debe garantizar que los participantes tengan la capacidad de tomar decisiones informadas y se respete su autonomía en el proceso de toma de decisiones.

8. Conclusiones

El cribado, ya sea oportunista o programado, es una estrategia de prevención secundaria que debe ser implementada con criterios rigurosos basados en las mejores pruebas científicas disponibles. El objetivo primordial del cribado es mejorar la salud de la población, lo cual requiere una evaluación constante de su beneficio-riesgo y de su coste-beneficio a lo largo de todo el proceso, desde la planificación hasta la implementación y el seguimiento del programa.

Es fundamental tener en cuenta tanto los aspectos organizativos como los éticos al llevar a cabo la implantación de un programa de cribado. Esto implica considerar la capacidad de recursos humanos y materiales necesarios, así como establecer protocolos claros para garantizar la calidad y eficacia del cribado. Asimismo, se deben respetar los principios éticos, como el consentimiento informado de los participantes y la equidad en el acceso al cribado.

Agradecimientos

A Beatriz Pérez Gómez por entregar altruistamente su ayuda y compartir su visión y su experiencia sobre el cribado.

9. Referencias

1. Celentano DD, Sklo M. Gordis Epidemiology. 6th edition. Elsevier, 2019. Tiene traducción al español: Celentano DD & Sklo M (autores) Gordis Epidemiología, 6ª edición. Barcelona: Elsevier Saunders. Septiembre 2019.
2. dos Santos Silva I. Cancer prevention. En Cancer epidemiology: principles and methods. Lión: International Agency for Research on Cancer, World Health Organization. 1999:355-83. Accesible en <https://publications.iarc.fr/Non-Series-Publications/Other-Non-Series-Publications/Cancer-Epidemiology-Principles-And-Methods-1999>. Consultado el 25 julio 2023.
3. García Aguado J. Cribar o no cribar: ¿esa es la cuestión? En: AEPap (ed.). Curso de Actualización Pediatría 2017. Madrid: Lúa Ediciones 3.0; 2017. p. 211-21.
4. Grupo de trabajo de la Ponencia de Cribado de la Comisión de Salud Pública. Documento marco sobre cribado poblacional. Diciembre 2010. Disponible en: https://www.sanidad.gob.es/areas/promocionPrevencion/cribado/docs/Cribado_poblacional.pdf. Consultado el 10 julio 2023.
5. Mariño MA. Incidentalomas: concept, relevance and challenges for medical practice. Revista Brasileira de Medicina de Família e Comunidade 10, n.º 35. 24 de junio de 2015.
6. Miller AB. Screening. En Ahrens W, Pigeot I (eds.) Handbook of epidemiology. Segunda edición. Nueva York: Springer. 2014:761-98.
7. Ministerio de Sanidad, Servicios Sociales e Igualdad. Informe del grupo de expertos sobre concreción de cartera común de servicios para cribado de cáncer de mama. Junio 2013. Disponible en: <https://www.sanidad.gob.es/areas/promocionPrevencion/cribado/docs/ResumenEjecutivoCribado-Cancer.pdf>. Consultado el 10 julio 2023.
8. Ministerio de Sanidad, Servicios Sociales e Igualdad. Resumen Ejecutivo del Cribado Neonatal Resumen ejecutivo del grupo de expertos sobre concreción de cartera común de servicios para cribado neonatal. Junio 2013. Disponible en: <https://www.sanidad.gob.es/areas/promocionPrevencion/cribado/docs/ResumenEjecutivoCribadoNeonatal.pdf>. Consultado el 10 julio 2023.
9. Molina-Barceló A, Moreno Salas J, Peiró-Pérez R, Arroyo G, Ibáñez Cabanell J, Vanaclocha Espí M, Binefa G, García M, Salas Trejo D. Desigualdades de acceso a los programas de cribado del cáncer en España y cómo reducirlas: datos de 2013 y 2020. Rev Esp Salud Publica 2021;95: e202101017.

10. Morrison AS. Screening. En Rothman KJ & Greenland S. Modern epidemiology. Segunda edición. Filadelfia: Lippincott Williams & Wilkins. 1998:499-518.
11. Newson AJ. Population screening. En Dawson A (ed.). Public health ethics. Key concepts and issues in policy and practice. Cambridge: Cambridge University Press. 2011:118-42.
12. Orozco-Beltrán D, Brotons Cuixart C, Banegas Banegas JR, Gil Guillén VF, Cebrián Cuenca AM, Martín Rioboó E, Jordá Baldó A, Vicuña J, Navarro Pérez J. Recomendaciones preventivas cardiovasculares. Actualización PAPPs 2022. Atención Primaria. 2022;54.
13. Piedrola Gil, Medicina preventiva y salud pública (10ª ed.) Ed. Masson, 2000.
14. Porta M. A dictionary of epidemiology. 6th ed. Oxford: Oxford University Press/International Epidemiological Association; 2014. 343 p. ISBN: 978-0-19-997673-7.
15. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299(17):926-930.
16. RGA Reinsurance Company. Screening for disease and disease risk: present and future considerations, caveats and consequences. Disponible en: https://www.rgare.com/docs/default-source/knowledge-center-articles/screening-for-disease-and-disease-risk_final1.pdf?sfvrsn=aa44e074_2. Consultado el 10 julio 2023.
17. Stevenson C. Screening, models of. En Gail MH & Benichou J (eds.). Encyclopedia of epidemiologic methods. Chichester: John Wiley & Sons, Ltd. 2000:817-39.
18. van Oortmarssen GJ. Screening, overview. En Gail MH & Benichou J (eds.). Encyclopedia of epidemiologic methods. Chichester: John Wiley & Sons, Ltd. 2000:839-47.
19. Wilson JMG & Jungner G. Principles and practice of screening for disease. WHO, Public Health Papers, no. 34, 1968 (http://apps.who.int/iris/bitstream/handle/10665/37650/WHO_PHP_34.pdf?sequence=17&isAllowed=y)

Anexo: El cribado en el cáncer: organismos de referencia y ejemplos de balance entre riesgos y beneficios

Es importante tener en cuenta que las recomendaciones de cribado del cáncer pueden variar según el tipo de cáncer, la población de pacientes y la disponibilidad de recursos de atención médica de cada región. A nivel internacional hay organismos de referencia en cuestión de recomendaciones de cribado:

- **La Organización Mundial de la Salud (OMS):** La OMS proporciona directrices y recomendaciones sobre el cribado del cáncer, así como información sobre las estrategias de prevención y control del cáncer a nivel global (<https://iris.who.int/handle/10665/275677>). Además, la Oficina Regional para Europa de la OMS, publicó “Screening programmes: a short guide” (<https://iris.who.int/bitstream/handle/10665/330829/9789289054782-eng.pdf?sequence=1>).
- **Agencia Internacional para la Investigación del Cáncer (IARC):** La IARC es parte de la OMS y se centra en la investigación sobre el cáncer. Publica informes sobre la eficacia y los riesgos asociados con diferentes métodos de cribado del cáncer (<https://www.iarc.who.int/>).
- **Instituto Nacional del Cáncer de los Estados Unidos (NCI):** El NCI es una agencia gubernamental de los Estados Unidos que proporciona información y directrices sobre el cribado del cáncer, así como recursos para pacientes y profesionales de la salud (<https://www.usa.gov/agencies/national-cancer-institute>).
- **Equipo de Trabajo de Servicios Preventivos de los Estados Unidos (USPSTF):** Este grupo emite recomendaciones basadas en la evidencia sobre el cribado y la prevención del cáncer, y sus recomendaciones a menudo influyen en las políticas de salud en los Estados Unidos (<https://www.uspreventiveservicestaskforce.org/uspstf/>). Además, recientemente, han desarrollado una aplicación móvil para ayudar a los médicos de atención primaria a identificar los servicios de detección, asesoramiento y medicación preventiva adecuados para sus pacientes. La información de la USPSTF se basa en sus recomendaciones más actualizadas y se puede buscar según algunas características específicas del paciente, como la edad, el sexo y estilos de vida o factores de riesgo modificables.

- **Grupo de trabajo de prevención del cáncer del programa de actividades preventivas y de promoción de la salud (PAPPS) de la Sociedad Española de Medicina Familiar y Comunitaria (semFYC).** Este grupo actualiza las pruebas científicas y emite recomendaciones basadas en la evidencia sobre prevención y detección precoz del cáncer en España, que son publicadas periódicamente en la revista *Atención Primaria* (<https://www.elsevier.es/es-revista-atencion-primaria-27-pdf-S0212656722001603>).

¿Se recomienda el cribado de cáncer de próstata?

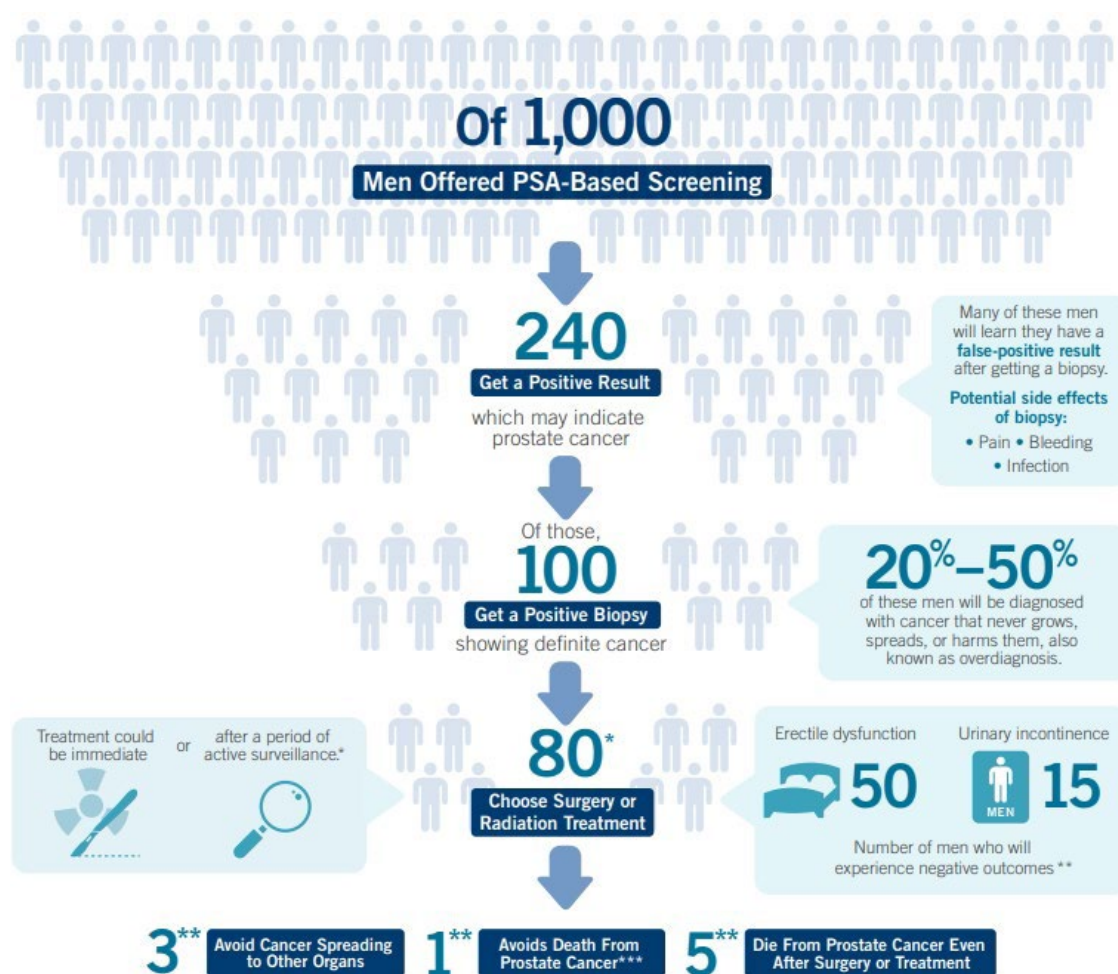
Según las recomendaciones del Equipo de Trabajo de Servicios Preventivos de Estados Unidos (USPSTF):

Para hombres de 55 a 69 años, la decisión de someterse a un cribado periódico basado en el antígeno específico de la próstata (PSA) para el cáncer de próstata debe ser individual. Para decidir si deben hacerse la prueba de cribado, deben tener la oportunidad de discutir con su médico los posibles beneficios y riesgos del cribado y considerar sus valores y preferencias en la decisión. El personal sanitario no debe someter a cribado a los hombres que no lo deseen.

Para el USPSTF, la recomendación para el cribado del cáncer de próstata en hombres de 55 a 69 años tiene una calificación de C, es decir, hay al menos una certeza moderada de que el cribado puede ofrecer un beneficio pequeño. En términos prácticos, la calificación de C supone ofrecer o proporcionar este servicio para pacientes seleccionados según sus circunstancias individuales (<https://sites.jamanetwork.com/spanish-patient-pages/2016/hoja-para-el-paciente-de-jama-160426.pdf>).

Sin embargo, para los hombres de 70 años o más, no se recomienda el cribado de rutina para el cáncer de próstata con PSA en este grupo de edad, ya que los riesgos de sobrediagnóstico y sobretratamiento superan los posibles beneficios.

Esta infografía creada por el USPSTF es muy ilustrativa:



Cribado de cáncer de mama

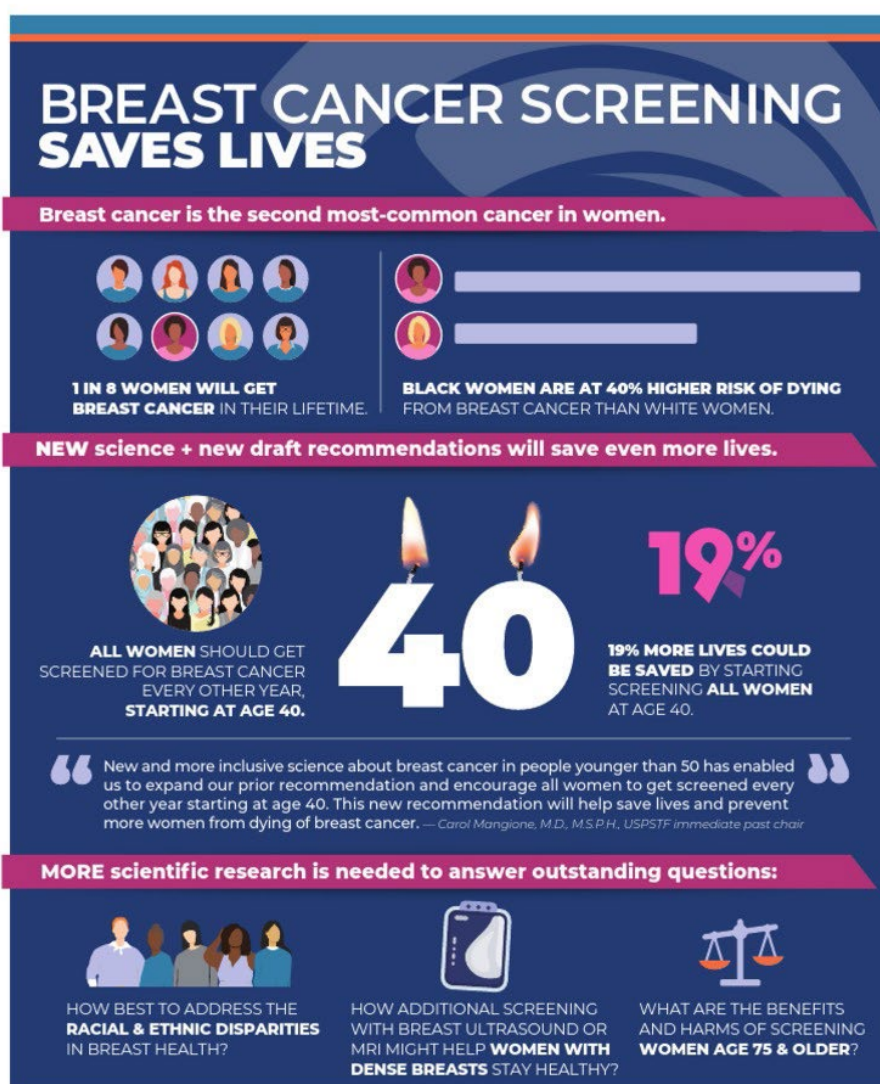
El Equipo de Trabajo de Servicios Preventivos de Estados Unidos (USPSTF) está actualizando las recomendaciones que emitió por última vez en 2016.

El USPSTF otorga una calificación de B para su recomendación de cribado bienal en el cáncer de mama en mujeres entre 50 y 74 años. La calificación de B significa que hay bastante certeza de que los beneficios netos son moderados o hay moderada certeza de que los beneficios netos son moderados o considerables. En términos prácticos, eso es equivalente a recomendar la provisión del servicio de cribado (<https://sites.jamanetwork.com/spanish-patient-pages/2016/hoja-para-el-paciente-de-jama-160426.pdf>). Para mujeres de 40 a 49 años, la decisión de someterse a un cribado con mamografía para el cáncer de mama debe ser individual (calificación de C); aunque en su actualización de 2023, actualmente en progreso, proponen adelantar a los 40 años la edad de inicio del cribado bienal del cáncer de mama, con una calificación de B. El PAPPs, en su última actualización

de 2022, no recomienda la mamografía de cribado en las mujeres de 40 a 49 años con riesgo medio de cáncer de mama (evidencia moderada, recomendación condicional en contra).

Por el contrario, en mujeres de 75 años o más, la recomendación del USPSTF tiene una calificación de I, es decir, no hay suficientes pruebas científicas para efectuar una recomendación concreta.

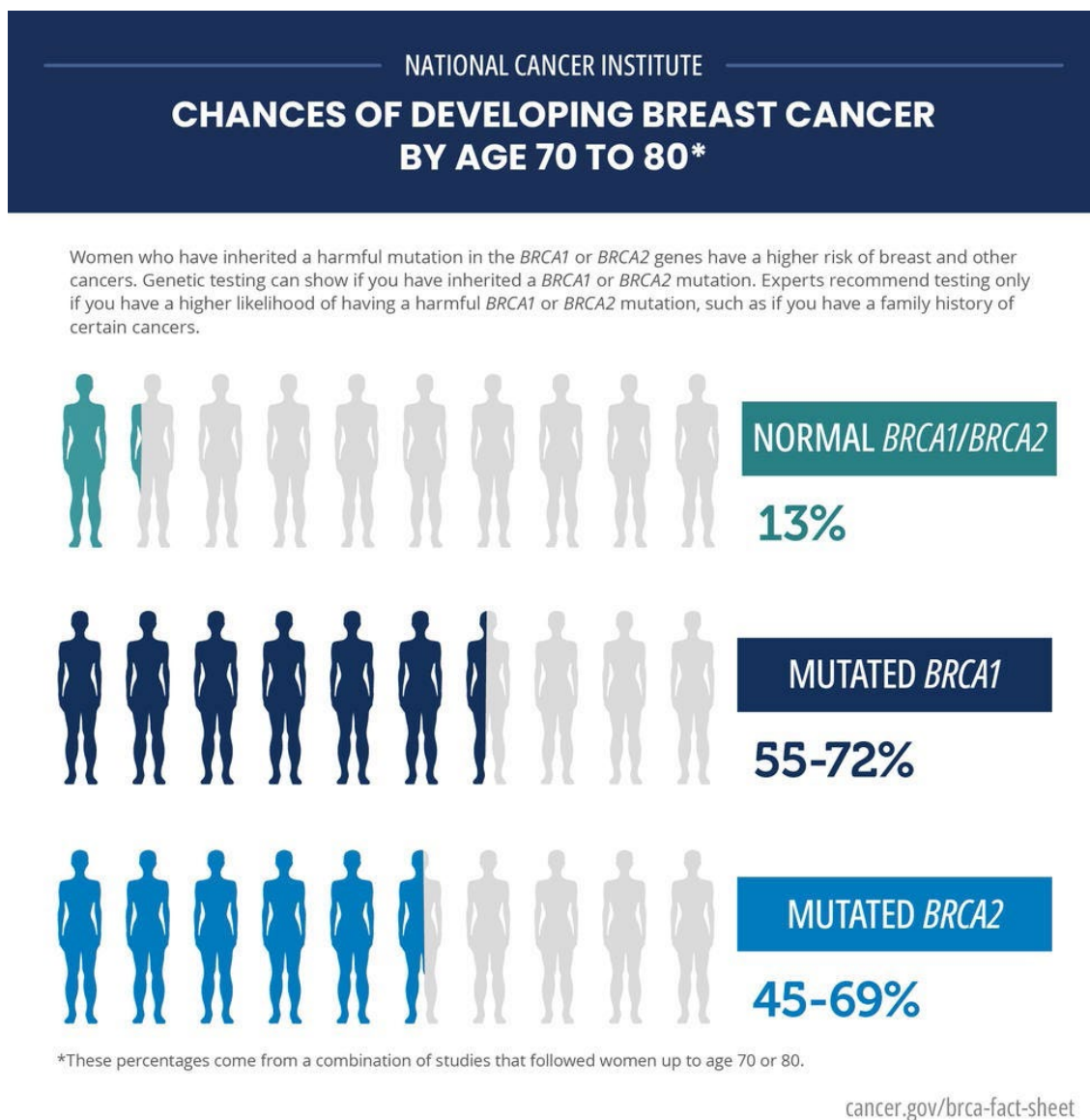
El USPSTF ha formulado recomendaciones sobre el uso de medicamentos para reducir el riesgo de cáncer de mama en mujeres, así como sobre la evaluación de riesgos, asesoramiento genético y pruebas genéticas relacionadas con BRCA1 o BRCA2 (incluyendo el cáncer de mama). Estas recomendaciones están disponibles en el sitio web del USPSTF (www.uspreventiveservicestaskforce.org), que además aporta esta infografía:



Join the conversation
and learn more about this draft recommendation.
www.uspreventiveservicestaskforce.org



Además, el *National Cancer Institute* (que forma parte de *National Institutes of Health*, en EEUU), ha realizado esta infografía para el riesgo de desarrollar cáncer de mama entre los 70 y 80 años:



Capítulo 15.

Revisiones sistemáticas y metanálisis

FERRÁN CATALÁ LÓPEZ

1. Introducción
2. Comenzar una revisión sistemática
3. Buscar los estudios relevantes
4. Seleccionar los estudios
5. Extraer la información de los estudios
6. Evaluar los estudios incluidos
7. Analizar los estudios incluidos. Metanálisis. Heterogeneidad
8. Evaluar metasesgos a través de los estudios
9. Interpretar y presentar los resultados de una revisión sistemática
10. Referencias

1. Introducción

“Se supone que la ciencia es acumulativa, pero los científicos rara vez acumulan el conocimiento científicamente.” (1)

Sir Ian Chalmers

En las últimas décadas, se ha producido una explosión en la generación de conocimiento. Se publican demasiados artículos en la literatura científica de cualquier campo para que sea posible mantenerse al día. Por ejemplo, anualmente se incluyen en MEDLINE® más de 1 369 000 nuevas publicaciones procedentes de algo más de 5200 revistas (2). Es poco habitual que se disponga del tiempo, los recursos y las habilidades para encontrar, recopilar, evaluar e interpretar todos los estudios relevantes sobre un tema, con el problema añadido de la proliferación de información errónea, el cóctel perfecto para la desinformación (3,4).

En medicina y salud pública, se reconoce desde hace tiempo la importancia de realizar síntesis rigurosas que permitan acumular el conocimiento disponible (5). Los resultados de un nuevo estudio deben interpretarse considerando, de manera rigurosa y actualizada, los resultados de otros estudios que abordan preguntas similares. Rara vez los resultados de un único estudio cambian completamente nuestra comprensión del conocimiento. Las revisiones sistemáticas permiten sintetizar el conocimiento disponible, aumentar la validez de las conclusiones de estudios individuales e identificar áreas donde es necesario realizar futuras investigaciones.

En este capítulo se presentan las bases conceptuales y metodológicas para la realización e interpretación de revisiones sistemáticas y metanálisis, y se discute su importancia en epidemiología y salud pública.

1.1. La revisión sistemática

Las revisiones sistemáticas son investigaciones científicas en las cuales la unidad de análisis son los estudios individuales o primarios. Son consideradas investigaciones sobre la propia investigación (“metainvestigación”) (6,7). Las revisiones sistemáticas tienen como objetivo reunir y sintetizar todo el conocimiento disponible sobre una pregunta claramente formulada, con métodos sistemáticos y explícitos para identificar, seleccionar y evaluar críticamente la investigación relevante, y para recopilar y analizar datos de los estudios primarios incluidos en la revisión (8,9). Cuando están bien diseñadas y realizadas, reducen los sesgos que podrían ocurrir en otros enfoques que revisan la literatura científica. En una revisión sistemática los métodos no sólo deben ser explícitos sino también sistemáticos, con

el propósito de producir resultados rigurosos y reproducibles, que proporcionen información útil para la toma de decisiones. Sin embargo, no todas las revisiones utilizan necesariamente métodos explícitos, rigurosos y reproducibles. Las revisiones “tradicionales” o narrativas no describen los métodos de manera transparente, examinan los resultados de sólo una parte de la investigación disponible y suelen presentarlos en función de las opiniones de los autores (10). En la Tabla 1 se presentan las principales etapas para la realización de una revisión sistemática.

Tabla 1: Etapas de una revisión sistemática

• Desarrollar la pregunta de la revisión. Por ejemplo, en formato PICOS: Población, Intervención, Comparador(es), Resultado (Outcome), Diseño del estudio (Study design)
• Elaborar un protocolo del estudio - Presentar los antecedentes y la justificación de la revisión - Definir los objetivos y los criterios de elegibilidad - Desarrollar la propuesta (o borrador) de estrategia de búsqueda - Identificar los métodos para evaluar la calidad (riesgo de sesgo) de los estudios - Describir los datos que serán extraídos - Predefinir las variables de resultado principales y los métodos de análisis - Registrar el protocolo en un repositorio de acceso público, como el registro PROSPERO, la plataforma <i>Open Science Framework</i> o algún repositorio institucional (este paso es opcional) - Publicar el protocolo en una revista con revisión por pares y en acceso abierto (por ejemplo, <i>Systematic Reviews</i> o <i>BMJ Open</i>) (este paso es opcional)
• Planificar la búsqueda de la literatura - Definir la(s) estrategia(s) de búsqueda y las fuentes de información - Con ayuda de expertos en documentación científica, revisar las búsquedas
• Identificar los estudios - Buscar en bases de datos electrónicas (por ejemplo, PubMed, EMBASE, Cochrane Library) - Utilizar otros métodos, si procede (por ejemplo, registros de estudios, búsquedas manuales de referencias bibliográficas, contactar a expertos...)
• Seleccionar los estudios - Prueba piloto de criterios de elegibilidad sobre una muestra - Cribado amplio de referencias a partir de título y resumen (recomendación: a realizar por dos personas revisoras de manera independiente) - Cribado estricto a partir del texto completo de los artículos (recomendación: a realizar por dos personas revisoras de manera independiente)
• Extraer los datos - Desarrollar una hoja de recogida de datos (por ejemplo, en Word, Excel o Google Forms) - Prueba piloto de la hoja de recogida de datos sobre una muestra de estudios incluidos - Extraer la información de interés incluyendo las variables de resultado principales y secundarias definidas en el protocolo (recomendación: a realizar por dos personas revisoras de manera independiente)
• Evaluar el riesgo de sesgo de los estudios incluidos - Utilizar una herramienta para evaluar la calidad (“riesgo de sesgo”) de los estudios (recomendación: a realizar por dos personas revisoras de manera independiente)

Tabla 1: Etapas de una revisión sistemática (continuación)

• Analizar los resultados - Evaluar la heterogeneidad clínica, metodológica y estadística - Sintetizar los resultados de manera cualitativa o cuantitativa (metanálisis), si procede - Realizar análisis adicionales (por ejemplo, análisis de sensibilidad, subgrupos, sesgos de publicación o por falta de resultados), si procede
• Presentar los resultados - Presentar los resultados de la selección de estudios (por ejemplo, diagrama de flujo) - Incluir una síntesis narrativa de los resultados principales, las características de las poblaciones y los estudios, y de los resultados de la evaluación de la calidad ("riesgo de sesgo") de los estudios - Presentar datos cuantitativos (por ejemplo, diagrama de efectos o <i>forest plot</i>) o cualitativos (por ejemplo, matriz temática, nubes de palabras) - Presentar los resultados de análisis adicionales, si procede
• Interpretar y discutir los resultados - Considerar la fortaleza (credibilidad) y aplicabilidad de los resultados - Discutir la relevancia de los resultados para los potenciales usuarios del conocimiento - Discutir las limitaciones (por ejemplo, de los estudios incluidos y de la propia revisión) - Proporcionar una interpretación general prudente a modo de conclusiones
• Informar sobre la disponibilidad de datos (extraídos y utilizados para todos los análisis), códigos de análisis, y cualquier otro material utilizado en la revisión, si procede
• Diseminar los resultados. Por ejemplo, publicación de un artículo científico, comunicación en congresos, medios de comunicación, redes sociales, etc.

1.2. Importancia de las revisiones sistemáticas

Las revisiones sistemáticas son herramientas esenciales para resumir el conocimiento de una manera rigurosa, creíble y precisa. Cada vez más, se utilizan para mantenerse actualizados sobre los avances que se producen dentro de una disciplina o un área de trabajo, por lo que se ha planteado que todas las personas que se inician en la investigación deberían empezar su etapa formativa elaborando al menos una revisión sistemática (11). Las revisiones sistemáticas proporcionan información para que los expertos elaboren recomendaciones y guías de práctica clínica; y los gestores de los servicios de salud formulen planes y políticas que consideren los riesgos, beneficios y costes de las tecnologías, medicamentos, intervenciones y programas de salud. Asimismo, son fundamentales para justificar, planificar y diseñar nuevos estudios que pretenden dar respuestas a incertidumbres que se plantean continuamente, proporcionando información para que las agencias financiadoras de la investigación valoren la necesidad y pertinencia de las nuevas investigaciones. Finalmente, pueden ayudar a resumir los resultados de la investigación para pacientes y sus familiares (8,9).

1.3. Tipos de revisiones sistemáticas

Cada año son incorporadas a MEDLINE® más de 8000 nuevas revisiones sistemáticas publicadas en inglés (12). La mayoría se centran en cuestiones relacionadas con la efectividad de las intervenciones sanitarias. Sin embargo, existen revisiones que persiguen otros objetivos, lo cual es un reflejo de la amplia variedad de preguntas y enfoques de investigación utilizados. En la Tabla 2 se presentan los tipos de revisiones sistemáticas más utilizados en epidemiología y salud pública (13-20).

Tabla 2: Tipos de revisiones sistemáticas

	Objetivo	Formato de pregunta	Ejemplo
Revisión de efectividad/seguridad de intervenciones sanitarias	Evaluar la efectividad/seguridad de una o varias intervenciones sanitarias en términos de su impacto en resultados de salud	PICOS – <u>P</u> oblación, <u>I</u> ntervención, <u>C</u> omparador, <u>R</u> esultado (<u>O</u> utcome), <u>D</u> iseño del estudio (<u>S</u> tudy design)	Intervenciones para prevenir caídas en personas mayores (13)
Revisión de evaluaciones económicas (eficiencia)	Evaluar la relación entre costes y resultados (coste-efectividad) de intervenciones, servicios o programas de salud	PICOC – <u>P</u> oblación, <u>I</u> ntervención, <u>C</u> omparador, <u>R</u> esultado (<u>O</u> utcome), <u>C</u> ontexto	Coste-efectividad de los tratamientos para cáncer infantil en países de ingresos medios y bajos (14)
Revisión de pruebas diagnósticas	Determinar la validez de una prueba diagnóstica en términos de su sensibilidad y especificidad para un diagnóstico concreto	PIRD – <u>P</u> oblación, <u>I</u> ndice, <u>p</u> ueba de <u>R</u> referencia, <u>D</u> iagnóstico de interés	Pruebas de anticuerpos para identificar infecciones pasadas o presentes por SARS-CoV-2 (15)
Revisión de factores pronósticos	Determinar el pronóstico de un problema de salud o enfermedad, la relación entre factores pronósticos específicos y un resultado en salud y/o modelos predictivos para el pronóstico	PFO – <u>P</u> oblación, <u>F</u> actor(es) pronóstico, <u>R</u> esultado (<u>O</u> utcome)	Modelos de predicción del riesgo de diabetes mellitus tipo 2 (16)
Revisión de etiología y/o factores de riesgo	Determinar la asociación entre una exposición/factor de riesgo y el resultado en salud	PEO – <u>P</u> oblación, <u>E</u> xposición, <u>R</u> esultado (<u>O</u> utcome)	Espacios verdes y su asociación con la mortalidad por todas las causas (17)
Revisión de prevalencia y/o incidencia	Determinar la prevalencia y/o incidencia de un problema de salud o enfermedad	PEC – <u>P</u> oblación, <u>E</u> nfermedad, <u>C</u> ontexto	Mortalidad y morbilidad (prevalencia) en personas sin hogar, presos, trabajadores sexuales y personas con trastornos por uso de sustancias (18)

Tabla 2: Tipos de revisiones sistemáticas (continuación)

	Objetivo	Formato de pregunta	Ejemplo
Revisión de experiencias u opiniones (cualitativa o métodos mixtos)	Investigar la experiencia, los valores o las opiniones de un fenómeno de interés	PIC – Población, Intervención o fenómeno de interés, Contexto	Valores y preferencias relacionadas con la salud y consumo de carne (19)
Revisión metodológica	Examinar aspectos metodológicos y su posible impacto en la calidad de la investigación	EDMO – Tipo de Estudios, Tipo de Datos, Tipo de Métodos, Resultado (Outcome)	Impacto de las búsquedas en registros de ensayos clínicos en revisiones sistemáticas de tratamientos farmacológicos (20)

En los últimos años han surgido otra clase de revisiones, con diferentes enfoques y metodologías, que también pueden ayudar a responder gran variedad de preguntas de investigación (21,22): las revisiones de alcance (“scoping reviews”) (23,24), las revisiones rápidas (“rapid reviews”) (25,26), las revisiones realistas (“realist reviews”) (27,28), las revisiones continuas o vivas (“living reviews”) (29,30), y las revisiones de revisiones (“overview of reviews”) (31,32). En la Tabla 3 se presentan sus principales características incluyendo algunos ejemplos (33-37).

Tabla 3: Otras clases de revisión

	Definición y características	Ejemplo
Revisión de alcance (<i>scoping review</i>)	Una revisión sobre un tema que identifica conceptos clave o el estado de la investigación en un área (a menudo de manera amplia y exploratoria). Puede incluir cualquier tipo de publicación y metodología de investigación. Suele plantearse como proyecto independiente o como paso preliminar para una revisión sistemática sobre una pregunta más concreta.	Utilidad de las redes sociales y datos de fuentes múltiples para actividades de farmacovigilancia (33)
Revisión rápida (<i>rapid review</i>)	Una revisión acelerada sobre una intervención, política o tema e intenta informar una decisión basada en pruebas científicas lo antes posible. Este tipo de revisiones sigue métodos simplificados para acortar los tiempos.	Eficacia de los equipos de protección personal para los trabajadores sanitarios que atienden a pacientes con enfermedad por filovirus (34)
Revisión realista (<i>realist review</i>)	Una revisión que produce teorías generalizables incorporando conceptos de investigación cualitativa, lo que implica considerar la complejidad de determinados fenómenos. Estas revisiones se enfocan en el componente contextual de las intervenciones y los mecanismos que llevan a un determinado resultado.	Comprender la eficacia de los programas de alimentación escolar (35)

Tabla 3: Otras clases de revisión (continuación)

	Definición y características	Ejemplo
Revisión continua o viva (<i>living review</i>)	Una revisión que se actualiza continuamente, incorporando nueva evidencia relevante a medida que está disponible. Si bien los métodos no son muy diferentes a otras revisiones sistemáticas, estas revisiones deben incluir además decisiones explícitas sobre con qué frecuencia se buscan y examinan nuevos estudios (por ejemplo, monitorización activa con búsquedas mensuales) y cómo son incorporados a la síntesis actualizada.	Tratamientos farmacológicos para COVID-19 (36)
Revisión de revisiones (<i>overview of reviews</i>)	Una revisión de múltiples revisiones sistemáticas sobre un problema o tema de salud específico, y que explora la coherencia de los métodos y resultados entre las revisiones. También se conocen como revisión paraguas (<i>umbrella review</i>).	Factores de riesgo ambientales en relación con la aparición de la esclerosis múltiple (37)

2. Comenzar una revisión sistemática

2.1. El equipo de revisión

Antes de comenzar una revisión sistemática, se debe identificar el equipo óptimo de investigación, determinado por el tipo de pregunta y formado por expertos con conocimientos sobre el tema y experiencia en revisiones, epidemiología y metodología de investigación (38). También es recomendable contar con el apoyo de un especialista en documentación científica (39) que ayude a diseñar, revisar y ejecutar las estrategias de búsqueda. A veces también se involucra en el proceso a usuarios finales (por ejemplo, responsables políticos, financiadores de la investigación, planificadores, gestores o pacientes).

2.2. Formular la pregunta y definir los criterios de elegibilidad

Plantear una pregunta clara y concisa es uno de los pasos más importantes de una revisión sistemática, ya que esta guiará todo el proceso de investigación (8,9). Las preguntas pueden tener un enfoque amplio o restringido, según los objetivos generales de la revisión. Cada modalidad de pregunta tiene sus ventajas e inconvenientes.

nientes. Por ejemplo, las preguntas amplias facilitan una evaluación más completa y aumentan la aplicabilidad de los resultados, pero requieren más recursos y resultan difíciles de interpretar.

Para formular la pregunta, se han propuesto distintos enfoques y formatos (Tabla 2). Por ejemplo, para revisiones sistemáticas que evalúan la efectividad o seguridad de las intervenciones sanitarias se establecen preguntas y criterios de elegibilidad de manera que se incluyan los componentes “PICOS” (ver un ejemplo en Tabla 4): la población (P); las intervenciones (I) o exposiciones; los comparadores o grupo control (C); los resultados de interés (O, del inglés *outcome*); y el diseño de los estudios (S, del inglés *study design*). Otros criterios de elegibilidad a considerar incluyen el estado de publicación (solo material publicado en revistas científicas o también material no publicado), el marco temporal y el idioma de las publicaciones (por ejemplo, solo artículos en inglés o también en otros idiomas). La pregunta y los criterios de elegibilidad deben justificarse y definirse adecuadamente para evitar ambigüedades durante la investigación, ya que ayudarán a definir las siguientes etapas de la revisión, comenzando por las búsquedas bibliográficas.

Tabla 4: Ejemplo de criterios de elegibilidad para una pregunta PICOS: ¿El tratamiento hipolipemiente con estatinas es beneficioso para reducir el riesgo cardiovascular?

Población	Intervenciones	Comparadores	Resultados (Outcomes)	Diseño de estudio
Adultos	Cualquier estatina (por ejemplo, simvastatina, atorvastatina o pravastatina)	Placebo	Mortalidad por todas las causas	Ensayos aleatorizados y controlados
Prevención primaria		No tratar	Mortalidad cardiovascular	Estudios observacionales
Prevención secundaria	Dosis	Otros tratamientos farmacológicos hipolipemiantes (por ejemplo, fibratos, resinas, ezetimiba)	Infarto agudo de miocardio	
	Duración del tratamiento	Intervenciones no farmacológicas	Ictus	
			Revascularizaciones	
			Presión arterial	

2.3. El protocolo de la revisión

El protocolo de una revisión sistemática describe de antemano el proceso y los métodos que se aplicarán (40), incluyendo detalles sobre la pregunta, los criterios de elegibilidad, las fuentes de información, las estrategias de búsqueda, la selección

de estudios, la extracción de información, la evaluación de los estudios, el análisis de datos y los planes de difusión. Especificar los métodos por adelantado reduce el riesgo de introducir sesgos en la revisión. Por ejemplo, unos criterios de inclusión claros evitan seleccionar estudios en función de sus resultados. Al contrario, con criterios de inclusión ambiguos o poco precisos se corre el riesgo de primar la inclusión de aquellos estudios cuyas conclusiones respalden la hipótesis de los investigadores.

El desarrollo de protocolos suele ser un proceso iterativo que requiere una comunicación adecuada dentro del equipo de revisión. Algunas instituciones, comités editoriales de revistas y autores abogan por el registro prospectivo de protocolos de revisiones sistemáticas para disminuir potenciales sesgos (por ejemplo, sesgos de publicación y de informe selectivo) y evitar duplicidades (41). Existe al menos un registro internacional prospectivo para protocolos de revisiones sistemáticas, PROSPERO (42,43), aunque hay otros repositorios donde se pueden registrar protocolos como el Open Science Framework (44). Además, algunos investigadores eligen publicar sus protocolos en revistas de acceso abierto (45).

3. Buscar los estudios relevantes

3.1. Fuentes de información

La pregunta y los criterios de elegibilidad de la revisión se utilizan para guiar las búsquedas de estudios relevantes en bases de datos electrónicas y otras fuentes complementarias (8,9,46). Las búsquedas para revisiones sistemáticas intentan ser extensas, con el fin de asegurar que se han incluido en la revisión tantos estudios relevantes como sea posible. Se recomienda utilizar más de una base de datos para mitigar posibles sesgos de indización y considerar el alcance de la revisión para seleccionar las más importantes. Algunas bases de datos electrónicas utilizadas habitualmente para localizar estudios de investigación en salud son MEDLINE®, EMBASE y las de la Cochrane Library. También se pueden utilizar otras de contenido específico (por ejemplo, CINAHL para investigación en cuidados de enfermería, PsycINFO en psicología y salud mental, SPORTDiscus en ciencias de la actividad física y del deporte y EconLit en economía) o geográficas (por ejemplo, IBECS —Índice Bibliográfico Español en Ciencias de la Salud y LILACS— Literatura Latinoamericana y del Caribe en Ciencias de la Salud).

A menudo se usan otras fuentes para complementar las bases de datos electrónicas: búsquedas manuales en revistas, en las listas de referencias bibliográficas de estudios incluidos o en registros de estudios (por ejemplo, [ClinicalTrials.gov](https://clinicaltrials.gov)). También se debería considerar la búsqueda de literatura gris o material difícil de localizar que puede incluir documentos alojados en repositorios institucionales (por ejemplo, informes, tesis doctorales), páginas web de agencias de financiación de la investigación, agencias reguladoras, industrias de medicamentos y tecnologías, o incluso repositorios de prepublicaciones (por ejemplo, [medRxiv.org](https://medrxiv.org)).

3.2. Estrategias de búsqueda

El desarrollo de las estrategias de búsqueda puede incluir diferentes enfoques para identificar palabras clave más o menos específicas, sus sinónimos y otros términos relacionados. Por ejemplo, en MEDLINE® se utiliza el índice de vocabulario controlado denominado MeSH (“Medical Subject Headings”). A veces es útil emplear filtros de búsqueda diseñados para recuperar registros específicos (47).

Las estrategias de búsqueda para cada una de las bases de datos, las fechas y períodos consultados, el número de registros recuperados y cualquier limitación de las búsquedas (por ejemplo, el idioma) deberán guardarse y presentarse en el informe o publicación de la revisión.

Es importante contar con la ayuda y consejo de alguna persona experimentada, especialista en documentación científica o bibliotecaria para guiar y mejorar las estrategias de búsqueda (por ejemplo, identificando errores o introduciendo mejoras en la selección de palabras clave y términos relacionados). La revisión por pares de las estrategias de búsqueda es un componente cada vez más valorado en las revisiones sistemáticas. Para ello, se puede utilizar la guía *Peer Review of Electronic Search Strategies (PRESS)* (48).

4. Seleccionar los estudios

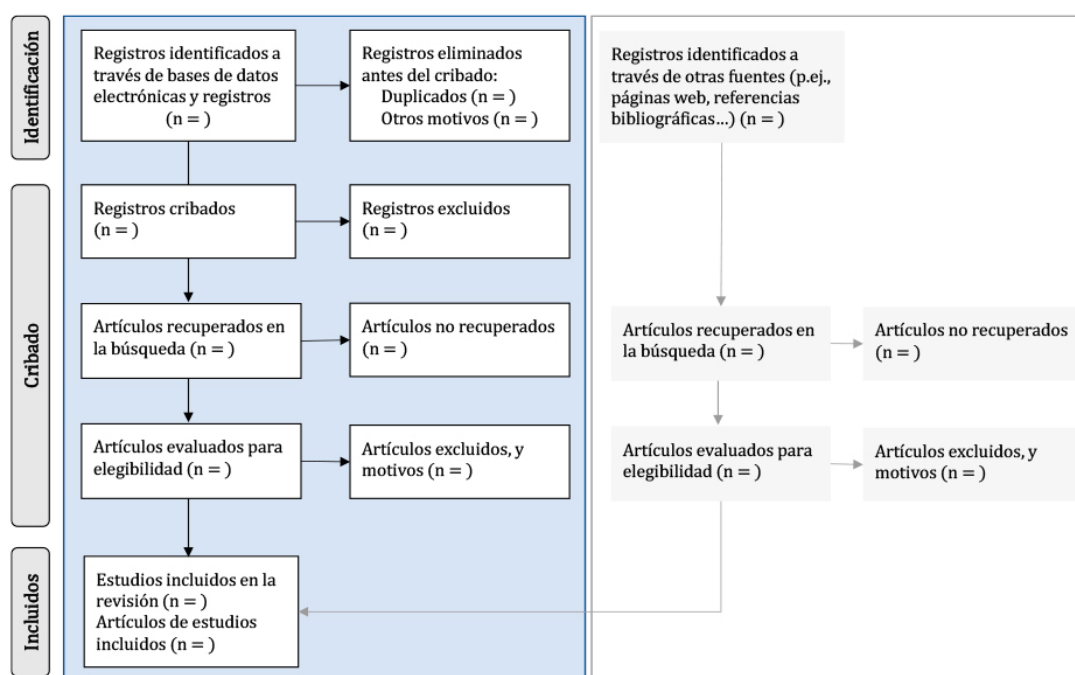
La selección de estudios es un proceso de varias etapas (8,9) en el que, tras eliminar los registros duplicados, el resto de los artículos identificados se criban mediante la lectura de títulos y resúmenes. A continuación, se lleva a cabo una revisión más profunda del artículo a texto completo. Cuando es necesario, se contacta con los investigadores del estudio para solicitar información adicional. Cada vez

más revisiones aplican una combinación de enfoques para seleccionar los estudios (por ejemplo, incorporando métodos automatizados para eliminar registros antes del cribado o priorizando registros durante el cribado) (49,50).

Para que no se excluyan estudios potencialmente relevantes, se recomienda que dos o más personas revisen y seleccionen los artículos de forma independiente tras realizar una prueba piloto con una pequeña muestra de registros que garantice la aplicación uniforme de los criterios de elegibilidad.

El proceso de selección de estudios requiere mantener un registro detallado de cada etapa ya que se debe informar todo el proceso de manera completa y transparente. Concretamente, en el apartado de resultados de la revisión se deberá presentar, idealmente con un diagrama de flujo (ver ejemplo en Figura 1), el número total de registros identificados (a partir de las búsquedas y otras fuentes), cribados e incluidos en la revisión, con los motivos de las exclusiones en cada etapa.

Figura 1: Ejemplo de diagrama de flujo.



5. Extraer la información de los estudios

Después de la selección, se debe extraer la información necesaria de las tablas, figuras y texto de las publicaciones de los estudios incluidos (8,9). En ocasiones las personas revisoras deben contactar a los investigadores de los estudios primarios para acceder a datos no publicados o solicitar información adicional.

Es aconsejable desarrollar una hoja o formulario de recogida de datos que incluya las variables a extraer: las características de los estudios (por ejemplo, países, ámbito, fuente de financiación), los elementos PICOS (por ejemplo, características demográficas y número de pacientes, enfermedad principal y comorbilidades, intervenciones y comparadores, variables de resultado o desenlace de interés, diseño del estudio, posibles factores de confusión cuando proceda) o los resultados del estudio (por ejemplo, número de acontecimientos o medidas resumen, estimaciones del efecto y su precisión). Realizar una prueba piloto sobre una muestra de estudios puede ayudar a garantizar que la extracción de datos es similar entre las personas revisoras.

En la revisión debe indicarse cuántas personas extrajeron la información de los estudios, si trabajaron de forma independiente o no (por ejemplo, datos extraídos por una persona y verificados por otra o datos extraídos por duplicado de manera independiente) y cualquier proceso utilizado para resolver desacuerdos entre revisores. Cuando dos o más personas extraen los datos de forma independiente disminuyen los errores en la extracción. También se debe indicar cualquier proceso utilizado para obtener o confirmar datos relevantes de los investigadores de los estudios primarios (por ejemplo, cómo fueron contactados, qué datos se solicitaron y si se obtuvo la información necesaria).

6. Evaluar los estudios incluidos

La validez de los resultados de una revisión sistemática dependerá en gran medida de la calidad de los estudios incluidos, determinada por el riesgo de sesgo. Ciertas características metodológicas de los estudios primarios pueden estar asociadas con el efecto de la intervención o con la exposición. Por ejemplo, los ensayos que describen métodos inadecuados de ocultación de la asignación pueden exagrar los efectos del tratamiento.

La revisión debe incluir una descripción del enfoque utilizado para evaluar el riesgo de sesgo, los constructos que se evalúan y una definición para cada uno, las opciones de juicio de las personas revisoras (por ejemplo, alto riesgo, bajo riesgo, riesgo incierto o con algunos problemas), el número de personas evaluadoras y el método de evaluación (por ejemplo, evaluación independiente y por duplicado) (8,9).

La evaluación puede realizarse usando escalas, listas de comprobación u otras herramientas disponibles que incluyen distintos componentes (o dominios específicos) del riesgo de sesgo. La Colaboración Cochrane ha desarrollado una herramienta para evaluar el riesgo de sesgo de los ensayos, denominada *Cochrane Risk of Bias* (RoB 2.0) *tool* (51). La *Newcastle-Ottawa Scale* (NOS) (52) se usa para estudios de cohorte y caso-control y la herramienta *Risk Of Bias In Non-Randomized Studies of Interventions* (ROBINS-I) (53) para estudios de intervención no aleatorizados. Independientemente de la herramienta utilizada, es preferible presentar evaluaciones para cada componente (o dominio) analizado en lugar de una única puntuación de calidad global, ya que permite comprender los elementos específicos que están en riesgo de sesgo para cada estudio (8).

La evaluación del riesgo de sesgo debe presentarse para cada uno de los estudios individuales o resultados de interés evaluados y puede incorporarse a los análisis para examinar si influye en los resultados de la revisión: dividir los estudios según el riesgo de sesgo en análisis de subgrupos o restringir el análisis principal a los estudios que se consideren de bajo riesgo en análisis de sensibilidad.

7. Analizar los estudios incluidos. Metanálisis. Heterogeneidad

En primer lugar, se debe determinar si los estudios se pueden analizar estadísticamente. La síntesis es el proceso de reunir la información de un conjunto de estudios con el objetivo de sacar conclusiones sobre el conocimiento disponible y depende de la pregunta de investigación y del tipo de información extraída. La síntesis puede ser narrativa o cuantitativa, que incluye análisis estadísticos (8,9).

Todas las revisiones sistemáticas incluirán al menos una síntesis narrativa que describa las características y los resultados de los estudios incluidos. Para ello, la información de los elementos PICOS se suele presentar en forma de tablas que permiten comparar las principales características de los estudios y su relevancia. En la Tabla 5 se presenta un ejemplo de las características de los estudios incluidos en una revisión sistemática sobre el uso de aspirina y el riesgo de fracturas (54).

Tabla 5. Características generales de los estudios incluidos

Autor (año)	Diseño	País	Número de pacientes	Edad media % mujeres	Seguimiento (meses)	Aspirina, identificación y dosis	Resultados evaluados
Bauer et al (1996)	Cohorte, prospectivo	Estados Unidos	7786	73,1-74,1 100	1,6	Autodeclarada 1-4 veces/ semana 5-7 veces/ semana	Fractura DMO
Bleicher et al (2011)	Transversal	Australia	1705	77,0 0	-	Comprobación en consulta -	DMO
Bonten et al (2017)	Transversal	Holanda	854	59,0 34	-	Comprobación en consulta 30-125 mg/día	Fractura DMO
Carbone et al (2003)	Transversal	Estados Unidos	2853	73,6 50	-	Comprobación en consulta 328 mg/día	Fractura DMO
Chuang et al (2016)	Caso-control	Taiwán	555	74,0 61	5	Historia pres- cripciones 106 mg	Fractura
Dobnig et al (2007)	Cohorte, prospectivo	Austria	1664	- 100	2	- -	Fractura
Hill et al (2008a)	Transversal	Trinidad y Tobago	340	63,9 100	-	Comprobación en consulta ≥ 3 veces/se- mana	DMO
Hill et al (2008b)	Transversal	Trinidad y Tobago	2501	56,3 0	-	Autodeclarada -	DMO
Lane et al (1997)	Transversal	Estados Unidos	4999	73,6 100	-	Autodeclarada 5-7 días/se- mana	DMO
Vestergaard et al (2006, 2012)	Caso-control	Dinamarca	498 617	43,4 52	1	Historia pres- cripciones ≤ 150 mg/día	Fractura
Vestergaard et al (2012)	Cohorte, prospectivo	Dinamarca	2016	50,8 100	10	Autodeclarada 325 mg/día	Fractura DMO
Williams et al (2011)	Caso-control anidado	Australia	1344	- 100	2	Comprobación en consulta -	Fractura

A partir de datos publicados por Barker et al. (54)

DMO: densidad mineral ósea.

En algunos casos, las revisiones incluyen además una síntesis cuantitativa que combina los resultados de los estudios aplicando análisis estadísticos (en adelante, metanálisis). El equipo de revisión deberá determinar si el metanálisis es factible y apropiado. Este paso implica determinar si los estudios incluidos son suficientemente similares (homogéneos) en cuanto a sus características para que la síntesis cuantitativa tenga sentido.

7.1. El metanálisis

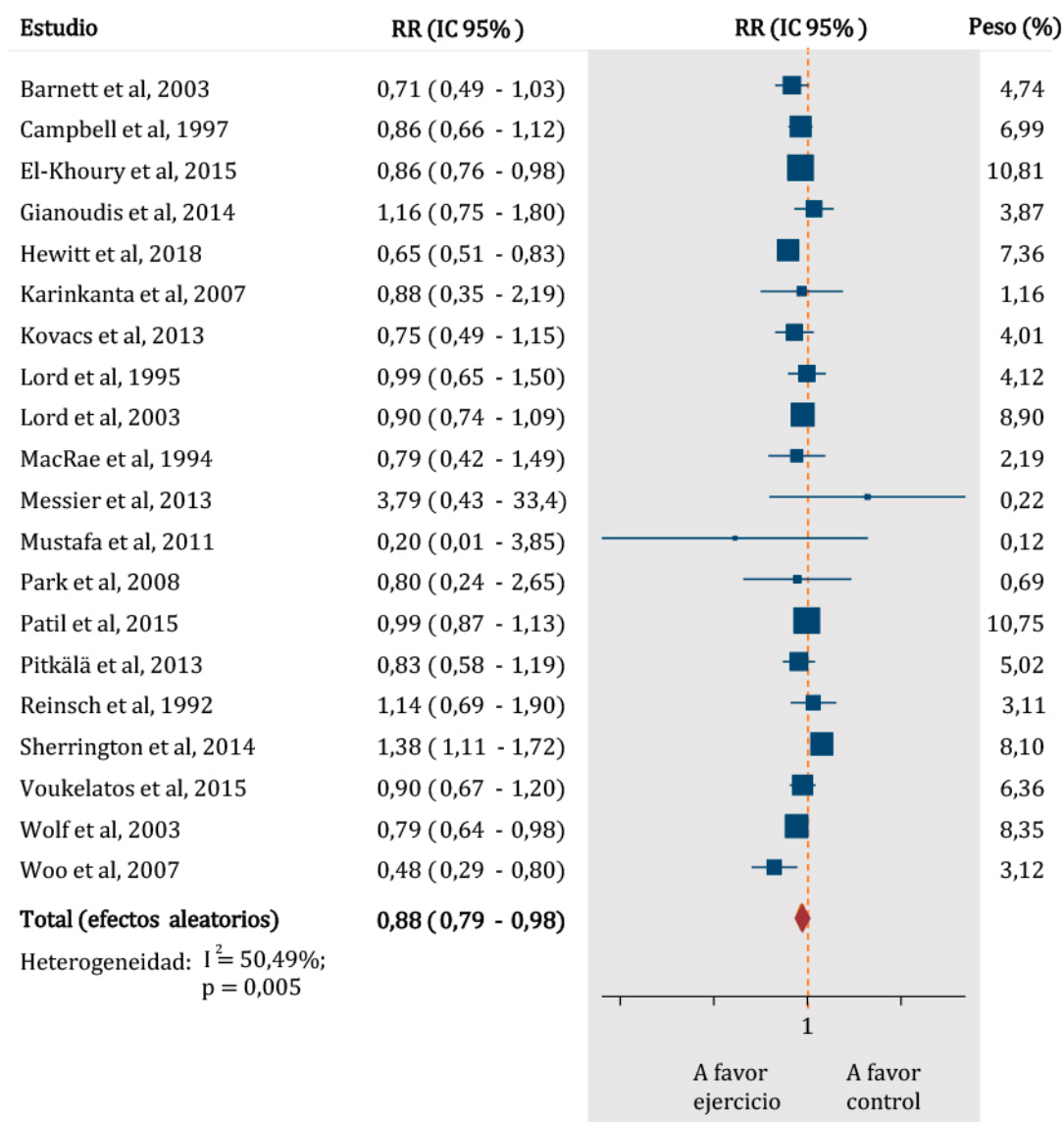
El metanálisis es la aplicación de técnicas estadísticas para resumir los resultados de dos o más estudios individuales incluidos en una revisión sistemática. El metanálisis mejora la precisión de las estimaciones, resuelve controversias que surgen de resultados contradictorios entre distintos estudios y permite responder preguntas no planteadas en los estudios individuales (8). Sin embargo, también puede presentar problemas si no se realiza y presenta de una manera adecuada (55,56).

La mayoría de los métodos de metanálisis calculan una media ponderada según la precisión de cada estudio, que depende sobre todo del tamaño muestral. Es importante reconocer el tipo de datos (por ejemplo, dicotómicos, continuos) en los resultados de los estudios individuales para seleccionar y calcular las medidas del efecto adecuadas para el análisis. Para un metanálisis de datos dicotómicos (por ejemplo, responder o no responder al tratamiento), se pueden calcular medidas como el riesgo relativo (RR), el odds ratio (OR) o la diferencia de riesgos; mientras que con datos continuos (por ejemplo, la presión arterial sistólica en mm Hg o puntuaciones en diferentes escalas de depresión) se pueden calcular la diferencia de medias, bruta o estandarizada. También se pueden combinar medidas del efecto previamente calculadas (por ejemplo, el logaritmo del OR ajustado en estudios observacionales).

En la Figura 2 se presentan los resultados de un metanálisis (57) de 20 ensayos que evaluaron el efecto de la práctica continuada de ejercicio sobre el riesgo de caídas en personas de 60 y más años. La representación gráfica de los resultados del metanálisis se hace a través de un diagrama de efectos (*forest plot*), donde en el eje de abscisas (eje X) se representa la medida del efecto de la intervención y en el eje de ordenadas (eje Y) se sitúan los diferentes estudios. Para cada estudio se muestra el efecto de la intervención de forma numérica y gráfica (RR y su intervalo de confianza) y los pesos relativos. En el gráfico, el tamaño del cuadrado que representa el estimador puntual del efecto es proporcional al tamaño de la muestra del estudio o a su precisión (por ejemplo, el inverso de la varianza). También se presenta la estimación del efecto combinado obtenido del metanálisis (en el gráfico, en forma de diamante). Los intervalos de confianza permiten comprobar si el efecto combi-

nado es estadísticamente significativo. En este ejemplo, el metanálisis estimó una reducción del riesgo de caídas del 12 % (RR = 0,88; intervalo de confianza del 95 % ([IC 95 %]) 0,79 – 0,98) con la práctica continuada de ejercicio.

Figura 2: Ejemplo de metanálisis. Efectividad de la práctica continuada de ejercicio y riesgo de caídas en personas mayores.



A partir de datos publicados por de Souto Barreto et al. (57). RR: riesgo relativo;
IC 95 %: intervalo de confianza del 95 %.

7.2. Modelos estadísticos en metanálisis

Existen dos modelos estadísticos para combinar los resultados en un metanálisis: el modelo de efectos fijos y el modelo de efectos aleatorios, cuya elección depende de la consideración de la hipótesis de homogeneidad del efecto entre los estudios.

El modelo de efectos fijos asume que existe un único y común efecto. Por tanto, la única fuente de variabilidad en los resultados procede de los estudios individuales y se cuantifica a través de la varianza. El modelo de efectos aleatorios asume que no hay un efecto común y único, sino una distribución de efectos, y que las diferencias observadas en los resultados entre los estudios también reflejan diferencias reales en esos efectos. Por tanto, además de la variabilidad de cada estudio (intraestudios), hay que contemplar la que se da entre los estudios, es decir, la desviación de cada estudio respecto al efecto promedio (τ^2). Ambos modelos ponderan cada estimación del efecto por el inverso de la varianza. Sin embargo, mientras que en el modelo de efectos fijos el peso (w) de cada estudio viene determinado exclusivamente por el inverso de la varianza (v) en cada estudio: $w_i = 1/v_i$, en el modelo de efectos aleatorios el peso (w) de cada estudio estará en función del inverso de la varianza (v) en cada estudio más la variabilidad entre los estudios (τ^2): $w_i = 1/(v_i + \tau^2)$. Existen diferentes métodos para asignar ponderaciones en ambos modelos (por ejemplo, el método de Woolf o del inverso de la varianza, el método de Mantel-Haenszel (58,59). Para el modelo de efectos aleatorios, también existen diferentes formas de estimar la variabilidad entre estudios (por ejemplo, el método de *DerSimonian-Laird* (60,61), el método de la máxima verosimilitud restringida (*Restricted Maximum-Likelihood* – REML) (62).

Independientemente de cuales sean los resultados del metanálisis, la interpretación es diferente según el modelo empleado. En el modelo de efectos fijos, los resultados se refieren a una población de estudios de características similares a las de los incluidos en la revisión; mientras que el modelo de efectos aleatorios permite generalizar los resultados a una población mayor de posibles estudios y obtener estimaciones más conservadoras, pero menos precisas, con intervalos de confianza más amplios. Cuando la variabilidad entre estudios (τ^2) es cero, los pesos de los estudios individuales serán los mismos para ambos modelos. En otras palabras, el modelo de efectos fijos es un caso particular de un modelo más general, el de efectos aleatorios, cuando no existe variabilidad entre estudios. Emplear un modelo u otro dependerá de los juicios que realice el equipo de revisión sobre las diferencias y similitudes (grado de homogeneidad) de los estudios que van a combinarse y de los supuestos mencionados anteriormente. En algunas revisiones sistemáticas publicadas, la elección del modelo está condicionada por alguna prueba estadística (55), pero se trata de una práctica desaconsejada (8).

7.3. Heterogeneidad

A cualquier tipo de variabilidad entre estudios se le puede llamar heterogeneidad. La variabilidad en las características de la población de estudio, las intervenciones y los resultados estudiados se describe como heterogeneidad clínica, que puede diferenciarse en: heterogeneidad anónima, desconocida o aleatoria, en la que los efectos observados difieren, pero se desconoce su causa, y heterogeneidad clasificada o conocida, que puede categorizarse y sus efectos ser reproducidos (63). La variabilidad en el diseño de los estudios y los riesgos de sesgos se describen como heterogeneidad metodológica. La variabilidad en los efectos observados entre los estudios se conoce como heterogeneidad estadística y puede ser consecuencia de las dos anteriores. Algunos autores reconocen que, puesto que todo metanálisis presenta heterogeneidad clínica y metodológica, la heterogeneidad estadística es inevitable (64).

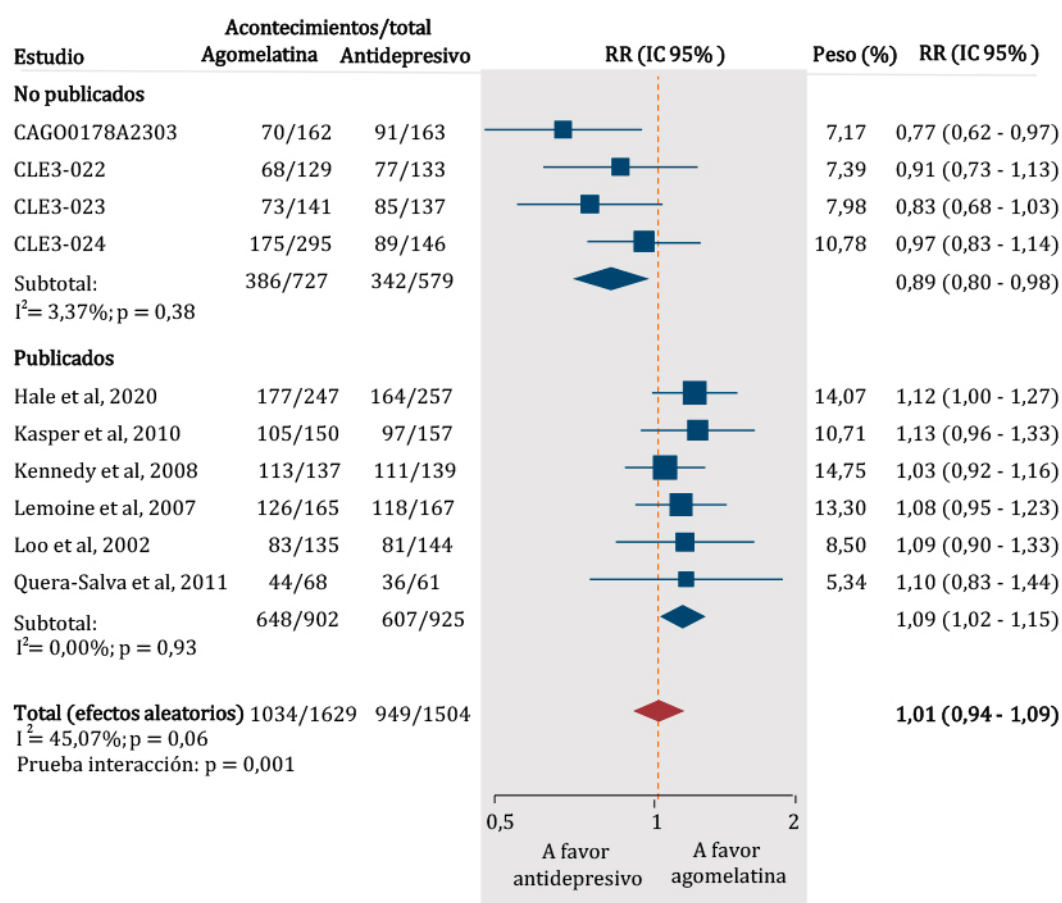
La heterogeneidad estadística puede examinarse visualizando el rango de las estimaciones de efectos y sus respectivos intervalos de confianza en el diagrama de efectos (*forest plot*). Si los intervalos de confianza para los resultados de los estudios no presentan solapamientos, es indicativo de heterogeneidad estadística. También existen pruebas estadísticas para identificar y medir la heterogeneidad: la Q de Cochran (65) y el índice I^2 (66). La Q de Cochran (65) es una prueba de ji cuadrado que evalúa si las diferencias observadas en los resultados son compatibles con el azar, comprobando la existencia de heterogeneidad bajo la hipótesis de que la variabilidad entre estudios es nula. Un valor de p bajo ($< 0,10$) es indicativo de heterogeneidad estadística. Sin embargo, la Q de Cochran tiene baja potencia estadística cuando el metanálisis incluye pocos estudios o estos tienen muestras pequeñas. Por tanto, aunque un resultado estadísticamente significativo indica heterogeneidad, un resultado no significativo no debe interpretarse como prueba de ausencia de heterogeneidad. El índice I^2 (66) indica la parte de la variabilidad total atribuible a la variabilidad entre estudios (heterogeneidad). Se expresa en valores porcentuales, considerando valores I^2 del 25 %, 50 % y 75 % como heterogeneidad baja, moderada y alta respectivamente. En el metanálisis de la Figura 2, el 50 % de la variabilidad en los efectos se debe a la variabilidad entre estudios.

7.4. Investigar las fuentes de heterogeneidad

La presencia de heterogeneidad estadística es habitual en los metanálisis publicados. Siempre que sea posible, hay que investigar las características de los estudios que puedan estar causando resultados diferentes. Los métodos más utilizados para investigar fuentes de heterogeneidad son los análisis de subgrupos, la metarregresión y los análisis de sensibilidad.

Los análisis de subgrupos estratifican los estudios para comparar los efectos en subgrupos con una prueba estadística de interacción. Los análisis de subgrupos analizan cómo los efectos en el metanálisis se asocian con características relacionadas con los pacientes (enfermedad leve frente a enfermedad grave y moderada), la intervención (dosis altas frente a dosis bajas de un tratamiento) o los estudios (estudios publicados frente a estudios no publicados). Por ejemplo, un metanálisis (67) de 10 ensayos que evaluaron la eficacia de agomelatina frente a otros antidepresivos en personas con depresión no mostró diferencias significativas entre grupos (RR = 1,01; IC 95 %: 0,94 – 1,09) con una heterogeneidad moderada ($I^2 = 46\%$). Los análisis de subgrupos mostraron que agomelatina resultó ser más eficaz que otros antidepresivos en los estudios publicados (RR = 1,09; IC 95 %: 1,02 – 1,15; $I^2 = 0\%$) y menos eficaz que otros antidepresivos en los estudios no publicados (RR = 0,89; IC 95 %: 0,80 – 0,98; $I^2 = 3\%$). Las diferencias entre los subgrupos resultaron ser estadísticamente significativas (p de interacción = 0,001) (*Figura 3*).

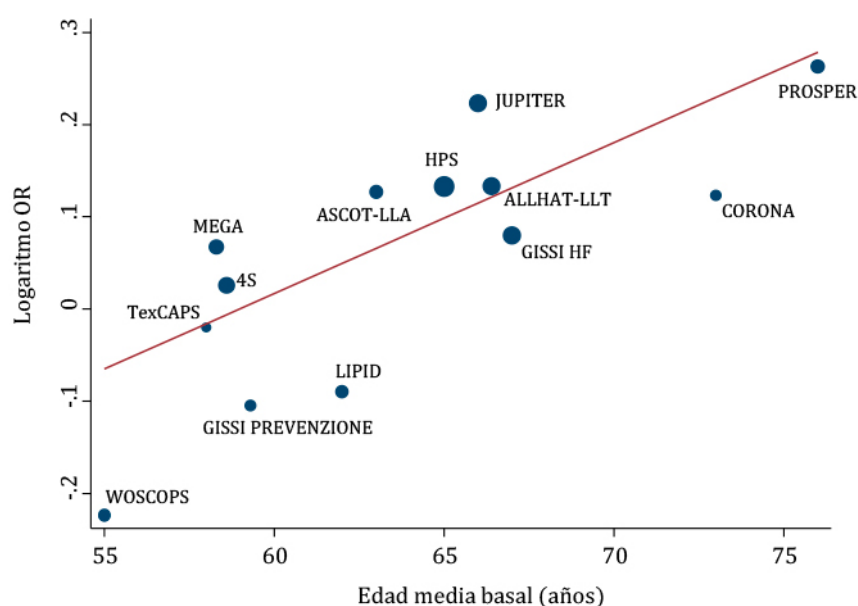
Figura 3: Análisis de subgrupos. Respuesta a agomelatina frente a otros antidepresivos.



A partir de datos publicados por Taylor et al. (67). RR: riesgo relativo; IC 95 %: intervalo de confianza del 95 %.

La metarregresión es una extensión del análisis de subgrupos que permite investigar el efecto de las características (variables continuas y categóricas) de los estudios, pero los análisis multivariantes son poco frecuentes, debido al número insuficiente de estudios (8,68). La metarregresión es conceptualmente similar a un modelo de regresión lineal tradicional, donde la unidad de análisis son los estudios individuales o primarios. Para reflejar la precisión de los resultados en los estudios, se aplican ponderaciones. Por tanto, la metarregresión es un modelo de regresión ponderada en el cual una variable de resultado (por ejemplo, estimación del efecto expresada como el logaritmo del OR) puede predecirse según los valores de una o más características de los estudios (variables independientes o explicativas), que suelen ser factores que pueden influir en el efecto de la intervención (por ejemplo, características basales de los pacientes: edad media, porcentaje de mujeres, porcentaje de comorbilidades, etc.). La metarregresión se debe utilizar solamente cuando se hayan incluido en el análisis al menos 10 estudios. Por ejemplo, un metanálisis (69) incluyó 13 ensayos en 91 140 pacientes, de los cuales 4278 (2226 recibieron estatinas y 2052 tratamiento de control) desarrollaron diabetes durante una media de 4 años. El tratamiento con estatinas se asoció con un incremento del 9 % en la incidencia de diabetes ($OR = 1,09$; $IC\ 95\ \% = 1,02 - 1,17$), con baja heterogeneidad ($I^2 = 11\ \%$). La metarregresión (Figura 4) mostró que el riesgo de diabetes fue más alto en los estudios que incluían pacientes más mayores ($\beta = 0,02$; $IC\ 95\ \% = 0,01 - 0,03$; $p = 0,004$; $I^2\ residual = 0\ \%$).

Figura 4: Metarregresión. Edad de los pacientes y riesgo de diabetes asociado al tratamiento con estatinas.



A partir de datos publicados por Sattar et al (69). OR: odds ratio.

En ocasiones se pueden realizar análisis de sensibilidad para examinar si los resultados del metanálisis son robustos frente a decisiones metodológicas, tales como la inclusión o exclusión de determinados estudios, la imputación por datos ausentes o la elección del modelo de análisis. Esto implica repetir los análisis utilizando decisiones diferentes a las que se tomaron originalmente. Los análisis de sensibilidad permiten realizar comparaciones informales entre diferentes formas de estimar el mismo efecto; mientras que los análisis de subgrupos realizan comparaciones estadísticas entre los mismos.

Todos estos métodos presentan limitaciones, se consideran exploratorios y, en el mejor de los casos, pueden ayudar a la generación de hipótesis. En la práctica puede suceder que haya pocos estudios incluidos en la revisión, que no se conozca qué características de estos influyen o que no se disponga de la información necesaria. Cada vez más, es habitual encontrar metanálisis que presentan un intervalo de predicción (70) que ayude a interpretar la heterogeneidad al estimar qué efectos se pueden esperar en nuevos estudios.

7.5. Desarrollos metodológicos en metanálisis

El desarrollo de técnicas de metanálisis es un campo dinámico y que cambia rápidamente. A continuación, se describen brevemente algunos desarrollos metodológicos.

- Metanálisis acumulado (*cumulative meta-analysis*) (71,72)

En un metanálisis acumulado los estudios se combinan secuencialmente, comenzando con un único estudio y agregando el resto de los estudios de uno en uno. De este modo, se muestra cómo cambia la estimación combinada a medida que se agrega cada estudio al análisis. Por ejemplo, un metanálisis acumulado (73) de estudios epidemiológicos observacionales realizados entre 1940 y 2002 evaluó el efecto de la posición de dormir (boca arriba o boca abajo) sobre el riesgo de muerte súbita del lactante. El metanálisis encontró un incremento del riesgo de muerte súbita del lactante asociado a dormir boca abajo, la posición recomendada durante décadas.

- Metanálisis de datos individuales de participantes (*individual participant data [IPD] meta-analysis*) (74)

En un metanálisis de datos individuales de participantes, en lugar de utilizar datos agregados de las publicaciones, se utilizan los datos originales de los estudios que son solicitados directamente a los investigadores. La mejor calidad de los datos y mayor abanico de posibles análisis producen

resultados más fiables. Por ejemplo, un metanálisis de datos individuales de pacientes (75) observó que el tratamiento con agentes estimulantes de la eritropoyesis incrementa el riesgo de mortalidad durante los períodos de tratamiento y empeora la supervivencia global en pacientes con cáncer.

- Metanálisis prospectivo (*prospective meta-analysis* [PMA]) (76)

Es un metanálisis de estudios que se identifican, evalúan y se determina su elegibilidad antes de que se conozcan sus resultados. Por ejemplo, varios equipos de investigación planificaron de manera prospectiva el desarrollo de ensayos similares para evaluar diferentes rangos objetivo para la saturación de oxígeno en bebés prematuros y combinar sus resultados (77). Aunque el metanálisis no encontró diferencias en el resultado primario compuesto de mortalidad o discapacidad mayor, se observó un incremento del riesgo de mortalidad para el rango objetivo de oxígeno más bajo.

- Metanálisis en red (*network meta-analysis* [NMA]) (78)

La mayoría de metanálisis se han centrado en la evaluación de los efectos de un tratamiento en comparación con un único comparador. El metanálisis en red permite evaluar de manera simultánea múltiples tratamientos que se no se han comparado directamente, pero de forma indirecta, a través de un comparador común (por ejemplo, placebo o estándar de tratamiento). Esta técnica permite obtener estimaciones de los efectos de cada tratamiento en relación con los demás de una manera eficiente. Por ejemplo, Cipriani *et al.* (79) estimaron la eficacia y aceptabilidad comparada de veintiún fármacos antidepresivos utilizados en el tratamiento agudo de la depresión mayor.

8. Evaluar metasesgos a través de los estudios

Las revisiones sistemáticas pretenden sintetizar todos los estudios relevantes sobre una pregunta de interés. Sin embargo, es posible que algunos estudios no se publiquen o que se publiquen de manera incompleta, inadecuada o selectiva (por ejemplo, según la magnitud, la dirección o la significación estadística de los resultados) (80,81). Si los estudios incluidos en una revisión sistemática proceden de una selección sesgada de toda la información que realmente existe, la validez de la revisión sistemática se verá seriamente afectada.

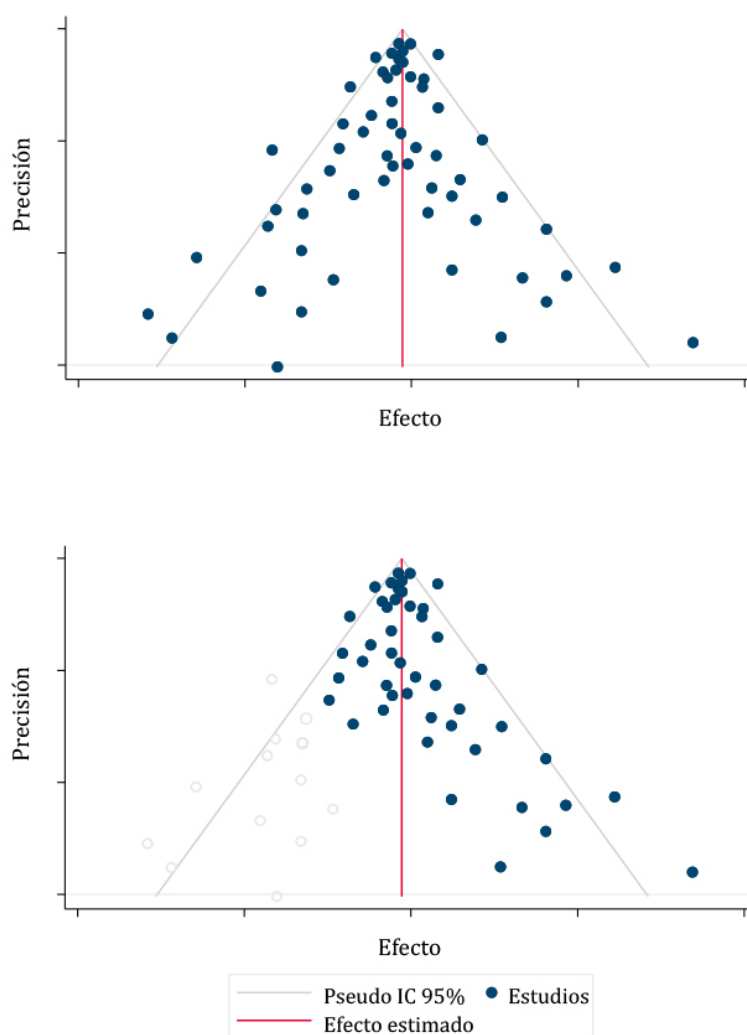
Los metasesgos o sesgos por falta de datos en la revisión sistemática se deben a la no publicación de estudios (sesgo de publicación) y a la presentación parcial de los resultados en los estudios incluidos (sesgo de informe selectivo) (8,81).

8.1. Evaluar la posibilidad del sesgo de publicación

En una revisión sistemática resulta difícil detectar o corregir el sesgo de publicación, pero se puede abordar, al menos parcialmente, intentando localizar todos los estudios relevantes publicados y no publicados, independientemente de sus resultados. Además, los resultados de los estudios incluidos en la revisión pueden proporcionar información sobre la falta de estudios.

Se han propuesto algunas pruebas gráficas para evaluar el sesgo de publicación. La principal es el uso de diagramas de embudo (*funnel plot*), que permite examinar la dispersión de las estimaciones de los efectos de la intervención en los estudios (por ejemplo, logaritmo del RR) en relación con su precisión (por ejemplo, el error estándar).

En la Figura 5 se presentan ejemplos de diagramas de embudo, con la medida del efecto de la intervención en el eje de abscisas (eje X) y la precisión en el eje de ordenadas (eje Y). Cuando hay un número suficiente de estudios (representados por puntos), y en ausencia de heterogeneidad y de sesgo de publicación, estos deberían adoptar el aspecto de un embudo, donde la parte más ancha corresponde a los estudios más pequeños (menos precisos). Si no faltan estudios, la forma del diagrama es simétrica. En cambio, si faltan estudios, no publicados por sus resultados nulos o contrarios a las hipótesis de los investigadores, la forma del diagrama es asimétrica. El sesgo de publicación puede deberse a la falta de estudios de pequeño tamaño que tienen estimaciones de efectos diferentes a las de los estudios más grandes (efectos de estudios pequeños). El diagrama de embudo es una técnica visual, subjetiva y de difícil interpretación en el caso de que existan pocos estudios (82), algo habitual en la mayoría de las revisiones sistemáticas publicadas (55).

Figura 5. Diagramas de embudo.

Nota: El diagrama de embudo situado en la parte superior presenta un ejemplo hipotético de gráfico simétrico en ausencia de sesgo. El diagrama de embudo situado en la parte inferior presenta un ejemplo hipotético de gráfico asimétrico en presencia de sesgo debido a la falta de estudios pequeños.

Es aconsejable complementar el diagrama de embudo con pruebas estadísticas. Las más utilizadas se basan en métodos de correlación (por ejemplo, prueba de Begg [83]) o regresión lineal (por ejemplo, prueba de Egger [84]), que contrastan la hipótesis nula de ausencia de sesgo de publicación. Para estas pruebas, un valor de p bajo (por ejemplo, $p < 0,10$) podría ser indicativo de sesgo de publicación (efecto de estudios pequeños).

Las pruebas para evaluar el sesgo de publicación deben utilizarse solamente cuando se han incluido en el metanálisis al menos 10 estudios, ya que con un número menor el poder estadístico de las pruebas es demasiado bajo. Todos estos métodos son indirectos y su idoneidad se basa en algunos supuestos estrictos que pueden tener difícil justificación (e interpretación) en la práctica (8,82).

8.2. Evaluar la posibilidad del sesgo de informe selectivo

Existen diversas aproximaciones para detectar el potencial sesgo de informe selectivo (40,80). Se pueden comparar los métodos descritos en el protocolo y los resultados presentados en el artículo publicado. Cuando el protocolo no está disponible, otra opción es comparar los apartados de métodos y resultados en el propio artículo publicado a la espera de que no presenten discrepancias. A veces, puede sospecharse que un resultado del que se dispone no ha sido publicado, como cuando se presentan resultados sobre efectos adversos de una intervención a partir de un umbral arbitrario (por ejemplo, únicamente aquellos con una frecuencia >10%).

Se han propuesto diversas herramientas, escalas y listas de comprobación para evaluar el riesgo del sesgo de informe selectivo. Sin embargo, la mayoría de los instrumentos disponibles presenta importantes limitaciones en términos de su alcance y las propiedades de medición (85). Recientemente, ha sido propuesto el primer enfoque estructurado para evaluar el riesgo del sesgo de informe selectivo en las revisiones sistemáticas, a través de la herramienta *Risk Of Bias due to Missing Evidence in a synthesis* (ROB-ME) (86).

9. Interpretar y presentar los resultados de una revisión sistemática

9.1. Interpretar correctamente los resultados de una revisión sistemática

La última parte de la revisión sistemática es la interpretación de los resultados obtenidos. Esta etapa debe incluir un resumen de los resultados principales, las limitaciones y los potenciales sesgos. Además, es importante que se discuta la fortaleza (credibilidad) y aplicabilidad de los resultados de la revisión. Para facilitar la interpretación de los resultados se puede utilizar el sistema *Grading of Recommendations Assessment, Development and Evaluation* (GRADE) (87), una herramienta que permite evaluar la calidad de los resultados y graduar la fuerza de las recomendaciones. También habría que considerar la relevancia de los resultados para usuarios finales ya que esto podría aumentar la transferencia del conocimiento.

La revisión sistemática deberá presentar una interpretación general (prudente) de los resultados considerando los objetivos, las limitaciones y el contexto de otras pruebas (o datos) relevantes. Las conclusiones pueden incluir recomendaciones es-

pecíficas para la toma de decisiones y las implicaciones para futuras investigaciones. Si no se pueden sacar conclusiones debido que el conocimiento disponible es insuficiente, esto se deberá indicar. Es muy importante considerar las limitaciones de los estudios incluidos (por ejemplo, riesgo de sesgos) y de la propia revisión sistemática (por ejemplo, sesgo de publicación y sesgo de informe selectivo). Por ejemplo, si la realización de los estudios incluidos es deficiente, las conclusiones de la revisión pueden estar sesgadas y esto deberá discutirse explícitamente. Además, las revisiones en sí mismas pueden presentar limitaciones. Para evitar sesgos, el equipo de revisión debe intentar incluir material no publicado relevante, hacer búsquedas manuales de material adicional, buscar en más de una base de datos electrónica, evaluar los sesgos de la propia revisión y actualizar periódicamente las búsquedas. A pesar de los esfuerzos para disminuir los potenciales sesgos en las revisiones sistemáticas, pueden persistir algunas limitaciones. Por ejemplo, la herramienta AMSTAR 2 (*A Measurement Tool to Assess systematic Reviews*) (88) puede ser útil para la lectura crítica e identificar limitaciones en las revisiones sistemáticas publicadas.

El último paso en la revisión es hacer que los resultados sean accesibles. Siguiendo prácticas de Ciencia Abierta (*Open Science*) (56,89,90), es recomendable informar sobre el acceso a los datos (extraídos y utilizados para todos los análisis) y cualquier otro material utilizado en la revisión. Finalmente, la forma más habitual de diseminar los resultados es la publicación a través de artículos en revistas científicas, aunque también puede hacerse a través de comunicaciones en congresos, redes sociales, medios de comunicación, etc.

9.2. Presentación clara y transparente de revisiones sistemáticas

Durante las últimas décadas, se han llevado a cabo importantes iniciativas para mejorar la transparencia y la calidad de la información presentada en las revisiones sistemáticas y metanálisis. Sin embargo, en ocasiones la presentación y descripción de algunas revisiones siguen sin ser del todo claras o falta información importante, fundamentalmente en los apartados de métodos y resultados (12,91).

En 2009 se publicó la declaración PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) (92), una guía de publicación que pretende facilitar la correcta presentación de los artículos que incluyen revisiones sistemáticas y metanálisis. Posteriormente, en marzo de 2021 fue publicada la declaración PRISMA 2020 (93-95), una actualización que reemplaza la versión anterior de la guía, a través de dos artículos: uno breve (94), que contiene la lista de comprobación y el diagrama de flujo en tres fases (identificación, cribado e inclusión), y otro más extenso (95) donde se detalla la explicación y la justificación de cada uno de los 27 elementos clave (ítems). También se explica el proceso de elaboración de la

propia guía, y se incluyen numerosos ejemplos de buena presentación a través de artículos publicados en la literatura.

La declaración PRISMA constituye una importante contribución a la estandarización metodológica de las revisiones sistemáticas y metanálisis. PRISMA 2020 pretende mejorar la claridad en la presentación de los métodos y resultados de los artículos científicos que sigan estas metodologías. Esta guía se ha convertido en una herramienta indispensable y útil para editores de revistas, revisores, autores y usuarios de revisiones sistemáticas y metanálisis. Para los metanálisis de estudios observacionales epidemiológicos se necesitan incluir algunos elementos ligeramente diferentes que se pueden encontrar en la guía *Meta-analysis of Observational Studies in Epidemiology* (MOOSE) (96). Se han desarrollado modificaciones y extensiones para otros tipos de revisiones sistemáticas y metanálisis (93). Por ejemplo, para las revisiones que utilizan datos individuales de participantes (PRISMA-IPD) (97), las revisiones que evalúan pruebas diagnósticas (PRISMA-DTA) (98) y aquellas que incorporan redes de múltiples tratamientos (PRISMA-NMA) (99). Del mismo modo, existen recomendaciones concretas para la presentación de protocolos de revisiones sistemáticas (PRISMA-P) (100), y para las búsquedas de la literatura en una revisión (PRISMA-S) (101). La mayoría de estas guías de publicación son requeridas por las principales revistas científicas que publican revisiones sistemáticas y metanálisis.

10. Referencias

1. Chalmers I, Hedges LV, Cooper H. A brief history of research synthesis. *Eval Health Prof.* 2002;25(1):12-37. doi: [10.1177/0163278702025001003](https://doi.org/10.1177/0163278702025001003)
2. US National Library of Medicine. MEDLINE PubMed Production Statistics, 2022 (consultado en febrero de 2023). Disponible en: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html
3. Scheufele DA, Krause NM. Science audiences, misinformation, and fake news. *Proc Natl Acad Sci U S A.* 2019;116(16):7662-7669. doi: [10.1073/pnas.1805871115](https://doi.org/10.1073/pnas.1805871115).
4. Ioannidis JPA, Stuart ME, Brownlee S, Strite SA. How to survive the medical misinformation mess. *Eur J Clin Invest.* 2017;47(11):795-802. doi: [10.1111/eci.12834](https://doi.org/10.1111/eci.12834).

5. Hunt M. How science takes stock: the story of meta-analysis. New York: Russell Sage Foundation; 1997.
6. Ioannidis JP, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and improvement of research methods and practices. PLoS Biol. 2015;13(10):e1002264. doi: [10.1371/journal.pbio.1002264](https://doi.org/10.1371/journal.pbio.1002264).
7. Ioannidis JPA. Meta-research: Why research on research matters. PLoS Biol. 2018;16(3):e2005468. doi: [10.1371/journal.pbio.2005468](https://doi.org/10.1371/journal.pbio.2005468).
8. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA. Cochrane Handbook for systematic reviews of interventions (Second Edition). Hoboken, NJ: Wiley-Blackwell; 2019.
9. Institute of Medicine Committee on Standards for Systematic Reviews of Comparative Effectiveness Research. Finding what works in health care: standards for systematic reviews. Washington: National Academies Press (US); 2011.
10. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126(5):376-80. doi: [10.7326/0003-4819-126-5-199703010-00006](https://doi.org/10.7326/0003-4819-126-5-199703010-00006).
11. Mahtani KR. All health researchers should begin their training by preparing at least one systematic review. J R Soc Med. 2016;109(7):264-8. doi: [10.1177/0141076816643954](https://doi.org/10.1177/0141076816643954).
12. Page MJ, Shamseer L, Altman DG, *et al.* Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med. 2016;13(5):e1002028. doi: [10.1371/journal.pmed.1002028](https://doi.org/10.1371/journal.pmed.1002028).
13. Tricco AC, Thomas SM, Veroniki AA, *et al.* Comparisons of interventions for preventing falls in older adults: a systematic review and meta-analysis. JAMA. 2017;318(17):1687-1699. doi: [10.1001/jama.2017.15006](https://doi.org/10.1001/jama.2017.15006).
14. Fung A, Horton S, Zabih V, Denburg A, Gupta S. Cost and cost-effectiveness of childhood cancer treatment in low-income and middle-income countries: a systematic review. BMJ Glob Health. 2019;4(5):e001825. doi: [10.1136/bmjgh-2019-001825](https://doi.org/10.1136/bmjgh-2019-001825).
15. Fox T, Geppert J, Dinnes J, *et al.* Antibody tests for identification of current and past infection with SARS-CoV-2. Cochrane Database Syst Rev. 2022;11(11):CD013652. doi: [10.1002/14651858.CD013652.pub2](https://doi.org/10.1002/14651858.CD013652.pub2).
16. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9:103. doi: [10.1186/1741-7015-9-103](https://doi.org/10.1186/1741-7015-9-103).

17. Rojas-Rueda D, Nieuwenhuijsen MJ, Gascon M, Perez-Leon D, Mudu P. Green spaces and mortality: a systematic review and meta-analysis of cohort studies. *Lancet Planet Health*. 2019;3(11):e469-e477. doi: [10.1016/S2542-5196\(19\)30215-3](https://doi.org/10.1016/S2542-5196(19)30215-3).
18. Aldridge RW, Story A, Hwang SW, *et al*. Morbidity and mortality in homeless individuals, prisoners, sex workers, and individuals with substance use disorders in high-income countries: a systematic review and meta-analysis. *Lancet*. 2018;391(10117):241-250. doi: [10.1016/S0140-6736\(17\)31869-X](https://doi.org/10.1016/S0140-6736(17)31869-X).
19. Valli C, Rabassa M, Johnston BC, *et al*. Health-related values and preferences regarding meat consumption: a mixed-methods systematic review. *Ann Intern Med*. 2019;171(10):742-755. doi: [10.7326/M19-1326](https://doi.org/10.7326/M19-1326).
20. Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. *BMJ*. 2017;356:j448. doi: [10.1136/bmj.j448](https://doi.org/10.1136/bmj.j448).
21. Moher D, Stewart L, Shekelle P. All in the Family: systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. *Syst Rev*. 2015;4:183. doi: [10.1186/s13643-015-0163-7](https://doi.org/10.1186/s13643-015-0163-7).
22. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Med Res Methodol*. 2018;18(1):5. doi: [10.1186/s12874-017-0468-4](https://doi.org/10.1186/s12874-017-0468-4).
23. Arksey H, O'Malley L. Scoping studies: Towards a Methodological Framework. *Int J Soc Res Methodol*. 2005;8(1):19-32. doi: [10.1080/1364557032000119616](https://doi.org/10.1080/1364557032000119616).
24. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*. 2018;18(1):143. doi: [10.1186/s12874-018-0611-x](https://doi.org/10.1186/s12874-018-0611-x).
25. Khangura S, Konnyu K, Cushman R, Grimshaw J, Moher D. Evidence summaries: the evolution of a rapid review approach. *Syst Rev*. 2012;1:10. doi: [10.1186/2046-4053-1-10](https://doi.org/10.1186/2046-4053-1-10).
26. Catalá-López F, Stevens A, Garritty C, Hutton B. Rapid reviews for evidence synthesis. *Med Clin (Barc)*. 2017;148(9):424-428. doi: [10.1016/j.medcli.2016.12.016](https://doi.org/10.1016/j.medcli.2016.12.016).

27. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review--a new method of systematic review designed for complex policy interventions. *J Health Serv Res Policy*. 2005;10 Suppl 1:21-34. doi: [10.1258/1355819054308530](https://doi.org/10.1258/1355819054308530).
28. Wong G, Greenhalgh T, Westhorp G, Buckingham J, Pawson R. RAMESES publication standards: realist syntheses. *BMC Med*. 2013;11:21. doi: [10.1186/1741-7015-11-21](https://doi.org/10.1186/1741-7015-11-21).
29. Elliott JH, Turner T, Clavisi O, Thomas J, Higgins JP, Mavergames C, Gruen RL. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med*. 2014;11(2):e1001603. doi: [10.1371/journal.pmed.1001603](https://doi.org/10.1371/journal.pmed.1001603).
30. Simmonds M, Elliott JH, Synnot A, Turner T. Living Systematic Reviews. *Methods Mol Biol*. 2022;2345:121-134. doi: [10.1007/978-1-0716-1566-9_7](https://doi.org/10.1007/978-1-0716-1566-9_7).
31. Gates M, Gates A, Pieper D, *et al*. Reporting guideline for overviews of reviews of healthcare interventions: development of the PRIOR statement. *BMJ*. 2022;378:e070849. doi: [10.1136/bmj-2022-070849](https://doi.org/10.1136/bmj-2022-070849).
32. Belbasis L, Bellou V, Ioannidis JPA. Conducting umbrella reviews. *BMJ Med*. 2022;1:e000071. doi: [10.1136/bmjmed-2021-000071](https://doi.org/10.1136/bmjmed-2021-000071).
33. Tricco AC, Zarin W, Lillie E, *et al*. Utility of social media and crowd-intelligence data for pharmacovigilance: a scoping review. *BMC Med Inform Decis Mak*. 2018;18(1):38. doi: [10.1186/s12911-018-0621-y](https://doi.org/10.1186/s12911-018-0621-y).
34. Hersi M, Stevens A, Quach P, *et al*. Effectiveness of personal protective equipment for healthcare workers caring for patients with filovirus disease: a rapid review. *PLoS One*. 2015;10(10):e0140290. doi: [10.1371/journal.pone.0140290](https://doi.org/10.1371/journal.pone.0140290).
35. Greenhalgh T, Kristjansson E, Robinson V. Realist review to understand the efficacy of school feeding programmes. *BMJ*. 2007;335(7625):858-61. doi: [10.1136/bmj.39359.525174.AD](https://doi.org/10.1136/bmj.39359.525174.AD).
36. Siemieniuk RA, Bartoszko JJ, Zeraatkar D, *et al*. Drug treatments for covid-19: living systematic review and network meta-analysis. *BMJ*. 2020;370:m2980. doi: [10.1136/bmj.m2980](https://doi.org/10.1136/bmj.m2980). Update in: *BMJ*. 2020;370:m3536. Update in: *BMJ*. 2020;371:m4852. Update in: *BMJ*. 2021;372:n858.
37. Belbasis L, Bellou V, Evangelou E, Ioannidis JP, Tzoulaki I. Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses. *Lancet Neurol*. 2015;14(3):263-73. doi: [10.1016/S1474-4422\(14\)70267-4](https://doi.org/10.1016/S1474-4422(14)70267-4).

38. Uttley L, Montgomery P. The influence of the team in conducting a systematic review. *Syst Rev*. 2017;6(1):149. doi: [10.1186/s13643-017-0548-x](https://doi.org/10.1186/s13643-017-0548-x).
39. Rethlefsen ML, Murad MH, Livingston EH. Engaging medical librarians to improve the quality of review articles. *JAMA*. 2014;312(10):999-1000. doi: [10.1001/jama.2014.9263](https://doi.org/10.1001/jama.2014.9263).
40. Moher D, Shamseer L, Clarke M, *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015;4(1):1. doi: [10.1186/2046-4053-4-1](https://doi.org/10.1186/2046-4053-4-1).
41. Stewart L, Moher D, Shekelle P. Why prospective registration of systematic reviews makes sense. *Syst Rev*. 2012;1:7. doi: [10.1186/2046-4053-1-7](https://doi.org/10.1186/2046-4053-1-7).
42. Booth A, Clarke M, Ghera D, Moher D, Petticrew M, Stewart L. An international registry of systematic-review protocols. *Lancet*. 2011;377(9760):108-109. doi: [10.1016/S0140-6736\(10\)60903-8](https://doi.org/10.1016/S0140-6736(10)60903-8).
43. Centre for Reviews and Dissemination. PROSPERO: International Prospective register of Systematic Reviews, 2023. <https://www.crd.york.ac.uk/prosperto/> (consultado en febrero de 2023).
44. Center for Open Science. Open Science Framework (OSF), 2023. <https://osf.io/> (consultado en marzo de 2023).
45. BioMed Central (BMC). Systematic Reviews, 2023. <https://systematicreviewsjournal.biomedcentral.com/> (consultado en febrero de 2023).
46. Lefebvre C, Glanville J, Wieland LS, Coles B, Weightman AL. Methodological developments in searching for studies for systematic reviews: past, present and future? *Syst Rev*. 2013;2:78. doi: [10.1186/2046-4053-2-78](https://doi.org/10.1186/2046-4053-2-78).
47. National Library of Medicine. Clinical Study Categories search filters. PubMed, 2023. <https://pubmed.ncbi.nlm.nih.gov/help/#clinical-study-category-filters> (consultado en febrero de 2023).
48. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol*. 2016;75:40-6. doi: [10.1016/j.jclinepi.2016.01.021](https://doi.org/10.1016/j.jclinepi.2016.01.021).
49. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. *J Clin Epidemiol*. 2022;144:22-42. doi: [10.1016/j.jclinepi.2021.12.005](https://doi.org/10.1016/j.jclinepi.2021.12.005).
50. O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Syst Rev*. 2019;8(1):143. doi: [10.1186/s13643-019-1062-0](https://doi.org/10.1186/s13643-019-1062-0).

51. Sterne JAC, Savović J, Page MJ, *et al.* RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. doi: [10.1136/bmj.l4898](https://doi.org/10.1136/bmj.l4898).
52. Wells GA, Shea B, O'Connell D, *et al.* The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: University of Ottawa; 2008. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp (consultado en febrero de 2023).
53. Sterne JA, Hernán MA, Reeves BC, *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919. doi: [10.1136/bmj.i4919](https://doi.org/10.1136/bmj.i4919).
54. Barker AL, Soh SE, Sanders KM, *et al.* Aspirin and fracture risk: a systematic review and exploratory meta-analysis of observational studies. *BMJ Open*. 2020;10(2):e026876. doi: [10.1136/bmjopen-2018-026876](https://doi.org/10.1136/bmjopen-2018-026876).
55. Page MJ, Altman DG, McKenzie JE, *et al.* Flaws in the application and interpretation of statistical analyses in systematic reviews of therapeutic interventions were common: a cross-sectional analysis. *J Clin Epidemiol*. 2018;95:7-18. doi:[10.1016/j.jclinepi.2017.11.022](https://doi.org/10.1016/j.jclinepi.2017.11.022).
56. Page MJ, Altman DG, Shamseer L, *et al.* Reproducible research practices are underused in systematic reviews of biomedical interventions. *J Clin Epidemiol*. 2018;94:8-18. doi:[10.1016/j.jclinepi.2017.10.017](https://doi.org/10.1016/j.jclinepi.2017.10.017).
57. de Souto Barreto P, Rolland Y, Vellas B, Maltais M. Association of Long-term Exercise Training With Risk of Falls, Fractures, Hospitalizations, and Mortality in Older Adults: A Systematic Review and Meta-analysis. *JAMA Intern Med*. 2019;179(3):394-405. doi: [10.1001/jamaintern-med.2018.5406](https://doi.org/10.1001/jamaintern-med.2018.5406).
58. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22(4):719-48.
59. Robins J, Greenland S, Breslow NE. A general estimator for the variance of the Mantel-Haenszel odds ratio. *Am J Epidemiol*. 1986;124(5):719-23. doi: [10.1093/oxfordjournals.aje.a114447](https://doi.org/10.1093/oxfordjournals.aje.a114447).
60. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. *Contemp Clin Trials*. 2015;45(Pt A):139-45. doi: [10.1016/j.cct.2015.09.002](https://doi.org/10.1016/j.cct.2015.09.002).
61. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-88. doi: [10.1016/0197-2456\(86\)90046-2](https://doi.org/10.1016/0197-2456(86)90046-2).
62. Veroniki AA, Jackson D, Viechtbauer W, *et al.* Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79. doi: [10.1002/jrsm.1164](https://doi.org/10.1002/jrsm.1164).

63. Catalá-López F, Tobías A. Metaanálisis de ensayos clínicos aleatorizados, heterogeneidad e intervalos de predicción. *Med Clin (Barc)*. 2014;142(6):270-4. doi: [10.1016/j.medcli.2013.06.013](https://doi.org/10.1016/j.medcli.2013.06.013).
64. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol*. 2008;37(5):1158-60. doi: [10.1093/ije/dyn204](https://doi.org/10.1093/ije/dyn204).
65. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954;10:101-29.
66. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-60. doi: [10.1136/bmj.327.7414.557](https://doi.org/10.1136/bmj.327.7414.557).
67. Taylor D, Sparshatt A, Varma S, Olofinjana O. Antidepressant efficacy of agomelatine: meta-analysis of published and unpublished studies. *BMJ*. 2014;348:g1888. doi: [10.1136/bmj.g1888](https://doi.org/10.1136/bmj.g1888).
68. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559-73. doi: [10.1002/sim.1187](https://doi.org/10.1002/sim.1187).
69. Sattar N, Preiss D, Murray HM, *et al*. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. *Lancet*. 2010;375(9716):735-42. doi: [10.1016/S0140-6736\(09\)61965-6](https://doi.org/10.1016/S0140-6736(09)61965-6).
70. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011;342:d549. doi: [10.1136/bmj.d549](https://doi.org/10.1136/bmj.d549).
71. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *J Clin Epidemiol*. 1995;48(1):45-57. doi: [10.1016/0895-4356\(94\)00106-z](https://doi.org/10.1016/0895-4356(94)00106-z).
72. Clarke M, Brice A, Chalmers I. Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. *PLoS One*. 2014;9(7):e102670. doi: [10.1371/journal.pone.0102670](https://doi.org/10.1371/journal.pone.0102670).
73. Gilbert R, Salanti G, Harden M, See S. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *Int J Epidemiol*. 2005;34(4):874-87. doi: [10.1093/ije/dyi088](https://doi.org/10.1093/ije/dyi088).
74. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:c221. doi: [10.1136/bmj.c221](https://doi.org/10.1136/bmj.c221).

75. Bohlius J, Schmidlin K, Brilant C, *et al.* Recombinant human erythropoiesis-stimulating agents and mortality in patients with cancer: a meta-analysis of randomised trials. *Lancet*. 2009;373(9674):1532-42. doi: [10.1016/S0140-6736\(09\)60502-X](https://doi.org/10.1016/S0140-6736(09)60502-X).
76. Seidler AL, Hunter KE, Cheyne S, Ghera D, Berlin JA, Askie L. A guide to prospective meta-analysis. *BMJ*. 2019;367:l5342. doi: [10.1136/bmj.l5342](https://doi.org/10.1136/bmj.l5342).
77. Askie LM, Darlow BA, Finer N, *et al.* Association between oxygen saturation targeting and death or disability in extremely preterm infants in the neonatal oxygenation prospective meta-analysis collaboration. *JAMA*. 2018;319(21):2190-2201. doi: [10.1001/jama.2018.5725](https://doi.org/10.1001/jama.2018.5725).
78. Riley RD, Jackson D, Salanti G, *et al.* Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ*. 2017;358:j3932. doi: [10.1136/bmj.j3932](https://doi.org/10.1136/bmj.j3932).
79. Cipriani A, Furukawa TA, Salanti G, *et al.* Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet*. 2018;391(10128):1357-1366. doi: [10.1016/S0140-6736\(17\)32802-7](https://doi.org/10.1016/S0140-6736(17)32802-7).
80. Chan AW, Song F, Vickers A, *et al.* Increasing value and reducing waste: addressing inaccessible research. *Lancet*. 2014;383(9913):257-66. doi: [10.1016/S0140-6736\(13\)62296-5](https://doi.org/10.1016/S0140-6736(13)62296-5).
81. Dwan K, Gamble C, Williamson PR, Kirkham JJ; Reporting Bias Group. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLoS One*. 2013;8(7):e66844. doi: [10.1371/journal.pone.0066844](https://doi.org/10.1371/journal.pone.0066844).
82. Sterne JA, Sutton AJ, Ioannidis JP, *et al.* Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*. 2011;343:d4002. doi: [10.1136/bmj.d4002](https://doi.org/10.1136/bmj.d4002).
83. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. 1994;50(4):1088-101. doi: [10.2307/2533446](https://doi.org/10.2307/2533446).
84. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-34. doi: [10.1136/bmj.315.7109.629](https://doi.org/10.1136/bmj.315.7109.629).

85. Page MJ, McKenzie JE, Higgins JPT. Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review. *BMJ Open*. 2018;8(3):e019703. doi: [10.1136/bmjopen-2017-019703](https://doi.org/10.1136/bmjopen-2017-019703).
86. Page MJ, Sterne JAC, Boutron I, *et al*. Risk Of Bias due to Missing Evidence (ROB-ME): a new tool for assessing risk of non-reporting biases in evidence syntheses, 2020. <https://www.riskofbias.info/welcome/rob-me-tool> (consultado en febrero de 2023).
87. Guyatt G, Oxman AD, Akl EA, *et al*. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64(4):383-94. doi: [10.1016/j.jclinepi.2010.04.026](https://doi.org/10.1016/j.jclinepi.2010.04.026).
88. Shea BJ, Reeves BC, Wells G, *et al*. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358:j4008. doi: [10.1136/bmj.j4008](https://doi.org/10.1136/bmj.j4008).
89. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. doi: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
90. Moreau D, Gamble B. Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychol Methods*. 2022;27(3):426-432. doi: [10.1037/met0000351](https://doi.org/10.1037/met0000351).
91. Nguyen PY, Kanukula R, McKenzie JE, *et al*. Changing patterns in reporting and sharing of review data in systematic reviews with meta-analysis of the effects of interventions: cross sectional meta-research study. *BMJ*. 2022;379:e072428. doi: [10.1136/bmj-2022-072428](https://doi.org/10.1136/bmj-2022-072428).
92. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097. doi: [10.1371/journal.pmed.1000097](https://doi.org/10.1371/journal.pmed.1000097).
93. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement, 2020. <http://www.prisma-statement.org/> (consultado en febrero de 2023).
94. Page MJ, McKenzie J, Bossuyt P, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71).
95. Page MJ, Moher D, Bossuyt P, *et al*. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*. 2021;372:n160. doi: [10.1136/bmj.n160](https://doi.org/10.1136/bmj.n160).

96. Stroup DE, Berlin JA, Morton SC, *et al.* Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283(15):2008-12. doi: [10.1001/jama.283.15.2008](https://doi.org/10.1001/jama.283.15.2008).
97. Stewart LA, Clarke M, Rovers M, *et al.* Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA. 2015;313(16):1657-65. doi: [10.1001/jama.2015.3656](https://doi.org/10.1001/jama.2015.3656).
98. McInnes MDF, Moher D, Thoms BD, *et al.* Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. 2018;319(4):388-396. doi: [10.1001/jama.2017.19163](https://doi.org/10.1001/jama.2017.19163).
99. Hutton B, Salanti G, Caldwell DM, *et al.* The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med. 2015;162(11):777-84. doi: [10.7326/M14-2385](https://doi.org/10.7326/M14-2385).
100. Shamseer L, Moher D, Clarke M, *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015;350:g7647. doi: [10.1136/bmj.g7647](https://doi.org/10.1136/bmj.g7647).
101. Rethlefsen ML, Kirtley S, Waffenschmidt S, *et al.* PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. doi: [10.1186/s13643-020-01542-z](https://doi.org/10.1186/s13643-020-01542-z).

Capítulo 16.

Planificación y ejecución de un cuestionario

ELENA PLANS BERISO Y FERNANDO GARCÍA LÓPEZ

- 1. Definición**
- 2. Selección de un cuestionario**
- 3. Validación de cuestionarios**
- 4. Diseño de las preguntas de un cuestionario**
- 5. Codificación**
- 6. Administración de cuestionarios**
- 7. Conclusiones**
- 8. Referencias**

1. Definición

Un cuestionario epidemiológico es un instrumento utilizado en la investigación epidemiológica para recolectar datos sobre la presencia, distribución y factores de riesgo de enfermedades o afecciones de salud en una determinada población; además, sirve para recabar otras informaciones útiles, como opiniones, sentimientos y rasgos de personalidad o carácter.

Los cuestionarios epidemiológicos tienen diversos usos en el campo de la salud. Algunos de ellos incluyen:

1. **La investigación de brotes de enfermedades.** Estos cuestionarios son utilizados para recopilar información detallada sobre los casos de enfermedad y los posibles factores de riesgo asociados. Esto ayuda a identificar la fuente de la infección y a controlar la propagación del brote.
2. **Estudios de prevalencia.** Los cuestionarios se emplean para estimar la prevalencia de una enfermedad o un fenómeno relacionado con la salud en una población específica. También permiten identificar sus factores asociados y determinar las necesidades de salud de dicha población.
3. **La investigación de factores de riesgo.** Los cuestionarios permiten recolectar información sobre los factores de riesgo de un proceso o una enfermedad, como la historia clínica, los antecedentes familiares, las exposiciones ambientales y los hábitos de vida. Esto ayuda a identificar las causas subyacentes y diseñar estrategias de prevención.
4. **La evaluación de programas de salud.** Los cuestionarios epidemiológicos son útiles para evaluar la efectividad de intervenciones y programas de salud pública. La recopilación de información antes y después de implementar las intervenciones permite determinar su impacto en la salud de la población.
5. **La vigilancia de enfermedades.** Los cuestionarios forman parte de los sistemas de vigilancia epidemiológica, permiten recopilar información continua sobre la incidencia y distribución de enfermedades en una población. Esto facilita la detección de tendencias y cambios en la salud de la población.

De este modo, los cuestionarios epidemiológicos son herramientas fundamentales en la investigación y la salud pública. Proporcionan datos valiosos para estudiar la salud de una población, identificar factores de riesgo y diseñar estrategias efectivas de prevención y control de enfermedades.

Encuesta y cuestionario

Encuesta y cuestionario son dos términos relacionados, pero con diferencias sutiles. Una encuesta se refiere a una recopilación de datos sistemática y estructurada realizada a través de una serie de preguntas para obtener información sobre un tema específico; mientras que un cuestionario es el instrumento utilizado en una encuesta para recopilar las respuestas de los participantes. En definitiva, una encuesta es un método general de recolección de datos y un cuestionario es el medio utilizado para llevar a cabo esa encuesta.

En los estudios epidemiológicos, los cuestionarios utilizados se denominan cuestionarios epidemiológicos. En el transcurso de un estudio epidemiológico, prácticamente cualquier recopilación de información lleva asociada un cuestionario, incluso en la recolección de muestras biológicas, en cuyo cuestionario se preguntaría al personal que trabaja en el estudio si la recogida de la muestra tuvo alguna incidencia. En algunos casos, es la única opción disponible y también puede resultar **menos costosa** que otros tipos de medidas, especialmente en el caso de cuestionarios autoadministrados.

Las encuestas nos permiten:

- Valorar estados **subjetivos**: opiniones, satisfacción, sentimientos, emociones, estados de ánimo, entre otros.
- Recopilar la **percepción** de los participantes sobre asuntos relacionados con la salud, más allá de lo que se puede abordar con instrumentos tecnológicos.
- Preguntar **hechos** relativos al participante o si ha habido algún problema en la obtención y conservación de una muestra biológica.

2. Selección de un cuestionario

2.1. Conocer los cuestionarios existentes

Antes de iniciar un estudio epidemiológico, es importante consultar si alguien ha realizado investigaciones similares. De la misma manera, cuando necesitamos un cuestionario, es conveniente comprobar si ya existen otros que puedan ser de utilidad en nuestro estudio: cuestionarios validados (ejemplo: el Cuestionario de Salud SF-36) o cuestionarios sin validar, contruidos *ad hoc* para un estudio previo.

Seleccionar un cuestionario preexistente permite comparar los resultados con los de estudios previos. Los validados son instrumentos de medida protocolizados y estandarizados, con normas estrictas y establecimiento de niveles de referencia. En los no validados, la experiencia nos proporciona información de qué preguntas funcionan mejor o peor.

Es crucial respetar completamente las instrucciones del cuestionario validado, evitando introducir nuevas preguntas basadas en criterios personales, ya que esto podría alterar completamente las respuestas en los demás ítems y comprometer así la validez del cuestionario.

Para encontrar cuestionarios y guías validados, podemos consultar bases de datos específicas (por ejemplo, BiblioPRO, ProQolid, FACIT, EORTC) o guías de escalas (por ejemplo, «La medida de la Salud. Guía de escalas de medición en español» de Xavier Badía y Jordi Alonso). Estas guías proporcionan información detallada para conocer los instrumentos disponibles y su nivel de desarrollo, lo que nos permite seleccionar el más adecuado tanto para la investigación como para la práctica clínica.

Existen diferentes tipos de cuestionarios para evaluar la calidad de vida relacionada con la salud. Por un lado, están los cuestionarios generales, como el Cuestionario SF-36, que se enfoca en evaluar la calidad de vida relacionada con la salud de manera más general. Por otro lado, existen cuestionarios específicos diseñados para pacientes con enfermedades particulares, por ejemplo, el cáncer. Estos últimos se adaptan a las necesidades y características de la enfermedad, con preguntas informativas y relevantes que permiten evaluar de manera precisa la calidad de vida de las personas que la padecen.

2.2. Valoración crítica de lo que existe

Una vez encontramos los cuestionarios existentes, hemos de tener varios aspectos en cuenta para elegir el que sea de mayor utilidad para nuestro estudio.

Lo ideal es que sea un cuestionario ampliamente utilizado y que esté validado, que garantice la utilidad, fiabilidad y validez de los datos recogidos. Los cuestionarios existentes creados *ad hoc* y no validados se pueden evaluar en el contexto del estudio mediante la reflexión sobre sus fortalezas y limitaciones y sobre si sería útil aplicar, al menos, parcialmente sus ítems.

2.3. Adaptación cultural

El proceso de adaptación cultural puede darse entre cuestionarios escritos en diferentes idiomas o, aunque estén en el mismo idioma, se apliquen en diferentes culturas. A continuación, nos vamos a centrar en el ejemplo de diferentes idiomas, siendo importante tener en cuenta que no se debe limitar a una simple traducción. Debe seguir una metodología rigurosa que asegure la equivalencia conceptual y semántica con el cuestionario original con el propósito de garantizar que sea comprensible y adecuado para la población objetivo. Recordemos que debemos estar seguros de que estamos midiendo lo que realmente queremos medir.

A continuación, se describen los pasos a seguir en esta fase de adaptación y validación del cuestionario en el idioma destino.

- a) **Traducción-retrotraducción.** Consiste en que al menos dos traductores de lengua nativa española y con amplio dominio del idioma original traduzcan al español el cuestionario y, posteriormente, otro traductor independiente, cuya lengua nativa sea el idioma de la versión original, y que desconozca el cuestionario original, haga la traducir inversa al idioma original. Si esta retrotraducción es equivalente a la versión original, la traducción se considera correcta, mientras que si no lo es habría que repetir el proceso de traducción-retrotraducción hasta que las versiones sean equivalentes.
- b) **Evaluación de contenido.** Se constituye un panel de expertos que revisa el cuestionario adaptado para asegurarse de que refleja de manera precisa las variables que se pretenden medir. Este panel busca garantizar que las preguntas sean relevantes y adecuadas para la población objetivo.
- c) **Preprueba.** Se realiza una prueba piloto del cuestionario adaptado en un grupo reducido de personas pertenecientes a la población objetivo. Durante esta fase, se busca identificar posibles problemas de comprensión, ambigüedades o dificultades que puedan surgir al responder el cuestionario. Se pueden llevar a cabo entrevistas o sesiones de grupo para recopilar comentarios y sugerencias de los participantes.
- d) **Análisis de resultados de la preprueba.** Los resultados de la preprueba se analizan para identificar las preguntas que generaron confusión o dificultades en la comprensión. Estas preguntas problemáticas se revisan y modifican para mejorar su claridad y adecuación al contexto cultural y lingüístico de la población objetivo.
- e) **Validación de constructo.** Se lleva a cabo un estudio más amplio con una muestra representativa de la población objetivo. En este estudio, se evalúa la concordancia interna de las preguntas y se realiza un análisis de validez

de constructo para verificar si las preguntas del cuestionario miden las variables que se pretenden medir.

- f) **Análisis de datos y refinamiento.** Los datos recolectados en el estudio de validación se analizan estadísticamente para evaluar la calidad psicométrica del cuestionario adaptado. Se realizan análisis de confiabilidad, validez convergente y discriminante, entre otros. En base a los resultados obtenidos, se refinan y ajustan las preguntas del cuestionario. Aquellas preguntas que no aportan información relevante o no son adecuadas para la población objetivo pueden ser eliminadas.

3. Validación de cuestionarios

Para que un cuestionario sea válido y útil no siempre es necesario realizar una validación formal, por ejemplo, en el caso de un cuestionario creado *ad hoc* para un estudio en particular. Sin embargo, hay situaciones en las que puede ser justificable validar cuestionarios ampliamente utilizados, en exposiciones ambientales o en el estudio de percepciones, por mencionar algunos ejemplos. En tales casos, se suele llevar a cabo un proceso sistemático de validación que está regulado por guías de comisiones internacionales de cuestionarios, escalas y test, y que suele implicar la publicación científica del proceso de validación.

Así, para validar un cuestionario, debemos aplicar los mismos criterios de validez y fiabilidad que exigimos a cualquier otro instrumento de medición, cuyas características principales se desarrollan a continuación.

3.1. Viabilidad

Debemos evaluar si el cuestionario puede ser utilizado con los recursos disponibles. Además, debemos tener en cuenta si es fácil de utilizar, contestar, puntuar y codificar.

3.2. Validez

En cuanto a la validez, es importante evaluar si el cuestionario mide realmente el fenómeno que se pretende medir. Esto puede ser más claro en fenómenos objetivos, como el peso, la altura o la presión arterial, donde existe una relación más clara entre lo que se observa y lo que representa¹.

Sin embargo, en los fenómenos subjetivos, como la ansiedad o la salud percibida, puede resultar más difícil evaluar si el cuestionario representa de manera exacta lo que se pretende medir. Algunos cuestionarios ya tienen un patrón de referencia establecido, es decir, escalas validadas, con las que compararlos.

Existen diferentes tipos de validez que se pueden considerar:

1. Validez de **presunción**. Para evaluar esta validez, se requiere la participación de expertos.
 - ♦ Validez **lógica** o **aparente** (*face validity*): evalúa si el cuestionario o ítem *parece* medir lo que pretende medir. Por ejemplo, si un test quiere medir si los estudiantes saben multiplicar y dividir y quienes lo juzgan piensan que el test es bueno para ello.
 - ♦ Validez de **contenido**: evalúa si el cuestionario aborda todos los aspectos o dimensiones relacionados con el concepto de estudio. Por ejemplo, un examen debería abordar contenidos que reflejen adecuadamente los conocimientos de la materia que se estudia.
2. Validez de **criterio**. Evalúa el grado de correlación entre la medida del cuestionario y la de un método de referencia.
 - ♦ Validez **concurrente**: compara la nueva medida con el patrón de referencia de modo simultáneo. Se aplica cuando se quiere desarrollar un cuestionario más breve y sencillo, como cuando se validaron los cuestionarios SF-12 y SF-20 con respecto al SF-36.

¹ Aunque la relación entre el instrumento de medida y los fenómenos objetivos sea relativamente clara, los instrumentos de medida no dejan de ser constructos relativamente complejos que hacen que las medidas no siempre correspondan con lo que se quiere medir. Por ejemplo, en la medición de la presión arterial hay múltiples factores que influyen en la validez de la medida, como la calibración del aparato, la colocación correcta del manguito en el brazo, la elección del brazo de la toma, la idoneidad del manguito en personas obesas o en niños, la posición del sujeto, que el sujeto esté en reposo durante unos minutos previos y no haya tomado café o haya fumado recientemente, la decisión de desechar o no la primera toma, la de efectuar dos o tres tomas, el lugar de la toma (el hospital, el centro de salud, el propio domicilio), incluso la presencia del llamado sesgo de dilución de la regresión a la media.

- ♦ Validez **predictiva**: la capacidad de una medida para predecir lo que sucederá un tiempo después en un patrón de referencia. Por ejemplo, la capacidad de un test efectuado antes del ingreso en la universidad para predecir si el alumno se graduará cuatro años después y con qué nota.
- 3. Validez de **constructo** o de **concepto**. Se evalúa cuando no existe un método de referencia disponible o no está al alcance del investigador. Se trata de una *teoría* o *constructo* que explica la relación entre varias conductas o actitudes.
 - ♦ Validez **convergente**: correlaciona los resultados con otras medidas, como se espera por la teoría. Por ejemplo, si queremos validar una escala de ansiedad y pensamos que las personas con más ansiedad tienen más síntomas relacionados con el sistema nervioso autónomo, la escala de ansiedad se correlacionaría con la de activación del sistema nervioso autónomo.
 - ♦ Validez **discriminante**: comprueba que la medida en estudio no se correlacione con otras medidas que la teoría no espera que se asocie. Por ejemplo, no esperaríamos encontrar una asociación entre la escala de ansiedad y un test de inteligencia si postulamos que son dos fenómenos independientes.

3.3. Fiabilidad o reproducibilidad

Debemos evaluar si los resultados del cuestionario son estables cuando se aplica en diferentes momentos o por diferentes entrevistadores.

Para determinar la fiabilidad de un cuestionario podemos evaluar la concordancia.

Concordancia interna

La concordancia interna (*internal consistency*) de un cuestionario es la coherencia en la respuesta a varios ítems del cuestionario relacionados con el mismo concepto. Por ejemplo, al evaluar el nivel de dependencia, los ítems relativos a actividades como vestirse, comer y asearse deben estar correlacionados (se suele usar el coeficiente *alfa* de Cronbach).

Concordancia externa

La concordancia externa es la estabilidad de los resultados con la repetición del cuestionario. Esto implica evaluar la concordancia **interobservador**, que mide

la concordancia entre las mediciones realizadas por dos observadores; la concordancia **intraobservador**, que mide la estabilidad de las mediciones realizadas por un mismo observador en momentos diferentes; y la concordancia **test-retest**, que mide la estabilidad de las respuestas dadas por un mismo sujeto al repetir el cuestionario (se suele usar el coeficiente *kappa* de Cohen).

Es importante tener en cuenta que hay limitaciones a la hora de valorar la fiabilidad de un cuestionario. Por un lado, si el tiempo entre la aplicación del cuestionario es muy largo, el fenómeno que se mide puede haber experimentado variaciones. Por otro lado, si el tiempo entre las aplicaciones es muy corto, puede haber un sesgo de recuerdo en las respuestas. Además, no todas las personas están dispuestas a completar el mismo cuestionario en dos ocasiones.

Existen factores que pueden afectar la concordancia o fiabilidad de un cuestionario, como las variaciones reales en el tiempo de la característica estudiada, los cambios en las condiciones de administración del cuestionario y las variaciones debidas al propio cuestionario, como preguntas formuladas de manera errónea o instrucciones poco claras y susceptibles de diferentes interpretaciones.

Cabe mencionar el **método del ancla**, que mejora tanto la concordancia interna como la externa. Establece puntos de referencia o *anclas* que ayudan a los participantes a comprender y evaluar diferentes elementos de una escala de medición. Consiste en proporcionar a los participantes una serie de declaraciones o ítems que representan diferentes niveles o extremos de la escala de interés. Por ejemplo, en un cuestionario de satisfacción con la salud, se pueden incluir ítems que representen niveles extremos como «totalmente insatisfecho» y «totalmente satisfecho». Estos ítems actúan como anclas y brindan a los participantes una referencia clara para evaluar su propio nivel de satisfacción. De esta manera, el método del ancla ayuda a reducir la ambigüedad y mejorar la concordancia en las respuestas de los participantes y facilita la comparación de las respuestas entre diferentes individuos al establecer una base común de evaluación.

3.4. Sensibilidad al cambio

La sensibilidad al cambio en los cuestionarios se refiere a la capacidad de detectar cambios significativos en la variable de interés a lo largo del tiempo o como resultado de una intervención. Un cuestionario sensible al cambio debe capturar tanto cambios sustanciales como pequeños. Se evalúa mediante métodos como el tamaño del efecto y la comparación de puntuaciones antes y después de la intervención. Es fundamental para evaluar los efectos de las exposiciones o de las intervenciones en los resultados del cuestionario.

El **tamaño de efecto** es una medida que indica la fuerza o la importancia de la relación o diferencia observada entre dos grupos o situaciones. En el contexto de los cuestionarios, se utiliza para cuantificar la magnitud del cambio experimentado por los participantes en una determinada variable de interés. Un tamaño de efecto más grande indica un cambio más sustancial. Existen diferentes formas de calcular el tamaño de efecto, como el coeficiente de correlación de Pearson, el coeficiente d de Cohen o la diferencia de medias estandarizada. Estas medidas permiten expresar el cambio en términos de desviaciones estándar y facilitan la comparación entre diferentes estudios o intervenciones.

4. Diseño de las preguntas de un cuestionario

El diseño de un cuestionario implica mucho más que simplemente formular preguntas. También implica tomar decisiones sobre qué cuestionarios preexistentes incluir, en caso de que se consideren necesarios o útiles, de modo que se aprovecha el conocimiento y la validación previa en el campo de investigación correspondiente, lo que puede proporcionar medidas fiables y válidas.

4.1. Tipo de información

En los cuestionarios se suelen recopilar diferentes tipos de información; cada tipo de información tiene su propia importancia y finalidad en el estudio y es crucial abordarlos con atención y respeto a la intimidad de los participantes. Algunos de los tipos de información más comunes son:

1. Información de **identificación**. Esta información se refiere a datos básicos que identifican a los participantes, como su nombre, dirección de correo electrónico, número de teléfono, etc. Estos datos son útiles para el registro y la comunicación con los participantes, así como para realizar un seguimiento adecuado del estudio.
2. Información **demográfica**. La información demográfica incluye características sociodemográficas de los participantes, como su edad, sexo, nivel de educación, ocupación, estado civil, sus ingresos, el lugar de residencia, entre otros. Estos datos proporcionan una descripción general de la población objetivo y permiten identificar posibles diferencias o tendencias de

los fenómenos relacionados con la salud en función de las características sociodemográficas.

3. Información **clínica**. En algunos estudios, es necesario recopilar información clínica sobre los participantes, como antecedentes médicos, diagnósticos, tratamientos previos, medicamentos en uso, historia familiar de enfermedades, entre otros. Esta información es relevante para conocer el estado de salud de los participantes y su relación con las variables de interés del estudio.
4. Información sobre la **exposición o factores de riesgo**. Estos tipos de preguntas están diseñados para recopilar información sobre la exposición de los participantes a ciertos factores de riesgo o variables de interés. Por ejemplo, en un estudio sobre enfermedades cardiovasculares, se pueden incluir preguntas relacionadas con el consumo de tabaco, la actividad física, la alimentación, los antecedentes de enfermedades cardiovasculares en la familia, entre otros.
5. Preguntas **delicadas**. Algunas preguntas en los cuestionarios pueden ser consideradas delicadas, ya que abordan temas personales o sensibles, como la salud mental, la actividad sexual, el consumo de drogas, etc. Estas preguntas requieren un enfoque cuidadoso para preservar la intimidad y confidencialidad de los participantes. Además, es importante brindar opciones de respuesta claras y respetar el deseo de los participantes de responder o no a estas preguntas.

4.2. Estructura de las preguntas

Las preguntas, según su estructura, pueden ser abiertas, cerradas o semicerradas.

Las preguntas **cerradas** son aquellas que ofrecen opciones predefinidas de respuesta, mientras que las preguntas abiertas permiten a los encuestados responder con sus propias palabras.

Para utilizar preguntas cerradas de manera efectiva, es importante tener en cuenta algunas consideraciones. Es recomendable evitar incluir demasiadas opciones (máximo de 10), ya que esto podría dificultar que los encuestados las recuerden. Además, es aconsejable utilizar alternativas mutuamente excluyentes, es decir, opciones que no se solapan. También se sugiere ordenar las alternativas de manera lógica, ya sea de menor a mayor, de acuerdo con una escala de intensidad o en orden alfabético.

Las respuestas dicotómicas tipo sí/no son un ejemplo común de preguntas cerradas. Estas preguntas son muy sencillas y se utilizan con frecuencia en cues-

tionarios largos. A menudo, se incluye una tercera respuesta como «No sabe/No contesta». Sin embargo, la principal desventaja de este tipo de preguntas es que no permiten matizar la respuesta, ya que los encuestados deben elegir entre dos opciones sin posibilidad de ofrecer detalles adicionales.

Las preguntas **abiertas** permiten a los encuestados responder con sus propias palabras y son útiles para recoger información compleja, ya que los encuestados pueden expresar sus opiniones y experiencias de manera libre y detallada. Sin embargo, las preguntas abiertas también presentan algunas dificultades. Requieren más tiempo para responder, ya que los encuestados deben elaborar sus respuestas en lugar de simplemente seleccionar una opción predefinida. Además, pueden plantear dificultades para codificar y analizar las respuestas. Al no estar limitadas a opciones específicas, las respuestas pueden ser más variadas y difíciles de clasificar y comparar de manera sistemática.

Las preguntas **semicerradas** aportan opciones de respuesta cerradas y permiten la posibilidad de expresar libremente otras respuestas. Un ejemplo de pregunta semicerrada sería «¿Qué medio de transporte utiliza habitualmente para ir al trabajo?

1. a pie
2. en coche
3. en autobús
4. en moto
5. en tren
6. en bicicleta
7. otros (especificar) _____ ».

La elección entre preguntas cerradas, semicerradas y abiertas dependerá de los objetivos y las necesidades específicas del estudio.

4.3. Función de las preguntas

Las preguntas de un cuestionario pueden desempeñar diferentes funciones. Estas incluyen las preguntas introductorias, las preguntas de cambio de tema, las preguntas amortiguadoras y las preguntas filtro. A continuación, se desarrolla cada una de estas funciones.

1. Preguntas **introductorias**. Se utilizan para establecer confianza en la conexión inicial con los encuestados, motivarlos a continuar con el cuestionario y establecer el tono del mismo. Suelen ser de naturaleza más general y pueden incluir cuestiones sobre demografía o preguntas de calentamiento para que el encuestado se sienta cómodo y se involucre en la encuesta.

2. Preguntas de **cambio de tema**. Se utilizan para hacer la transición de un tema a otro dentro del cuestionario. Estas preguntas ayudan a cambiar el enfoque de la encuesta y guiar al encuestado hacia un nuevo conjunto de preguntas relacionadas con un tema diferente. Son especialmente útiles cuando se abordan diferentes áreas temáticas en un cuestionario y se requiere una transición suave entre ellas.
3. Preguntas **amortiguadoras**. Se utilizan para suavizar el impacto de las preguntas sensibles o delicadas que podrían resultar incómodas para los encuestados. Estas preguntas se colocan estratégicamente antes de las más sensibles para preparar al encuestado y reducir la resistencia o la evitación de responder. Por ejemplo, se puede comenzar con cuestiones generales antes de abordar temas más personales o íntimos para que el encuestado se sienta más cómodo y dispuesto a responder.
4. Preguntas **filtro**. Se utilizan para clasificar a los encuestados en diferentes grupos o segmentos con base en ciertas características o criterios. Estas preguntas se colocan al principio del cuestionario para determinar si el encuestado cumple con los requisitos o criterios necesarios para participar en la encuesta completa. Si no los cumple, se le puede dirigir a un flujo de preguntas alternativo o puede ser excluido del resto del cuestionario. También son útiles para asegurar que los datos recopilados sean relevantes para el objetivo de la encuesta y para evitar el desperdicio de recursos en encuestados no calificados.
5. Preguntas de **verificación**. Se utilizan para validar la exactitud y concordancia de las respuestas del encuestado. Se incluyen preguntas de verificación para asegurarse de que el encuestado esté prestando atención y respondiendo de manera precisa. Por ejemplo, se puede incluir una pregunta repetida con ligeras variaciones en diferentes secciones del cuestionario para verificar si las respuestas son coherentes.
6. Preguntas de **clasificación**. Se utilizan para segmentar a los encuestados en grupos o categorías específicas con base en características demográficas, preferencias o comportamientos. Estas preguntas permiten obtener información descriptiva sobre la muestra y pueden ayudar a realizar análisis más detallados y segmentados de los resultados.
7. Preguntas de **seguimiento**. Se utilizan para profundizar en las respuestas anteriores del encuestado y obtener más detalles o aclaraciones sobre ciertos aspectos. Estas preguntas permiten obtener información adicional y más precisa sobre temas específicos y pueden ayudar a enriquecer los datos recopilados.

8. Preguntas de **evaluación**. Se utilizan para evaluar la satisfacción del encuestado con respecto a un producto, servicio o experiencia específica. Estas preguntas permiten medir la opinión y la percepción del encuestado y son comúnmente utilizadas en estudios de satisfacción del cliente o evaluaciones de calidad.
9. Preguntas de **opinión**. Se utilizan para recopilar las opiniones y los puntos de vista del encuestado sobre un tema en particular. Estas preguntas permiten capturar la perspectiva individual y pueden ser útiles en estudios de mercado, encuestas políticas o investigaciones de opinión pública.
10. Preguntas de **intención o comportamiento futuro**. Se utilizan para obtener información sobre las intenciones o comportamientos futuros del encuestado en relación con un producto, servicio o actividad específica. Estas preguntas pueden proporcionar información valiosa sobre las tendencias y las preferencias futuras, y son útiles para la planificación y toma de decisiones.

4.4. Redacción de las preguntas

Para mejorar la calidad de los datos, se recomienda redactar las preguntas según las siguientes pautas:

1. Evitar preguntas **ambiguas**. Formular las preguntas de manera clara y precisa, evitando ambigüedades que puedan llevar a respuestas confusas o incorrectas. Se debe ser específico y evitar términos o frases que puedan tener múltiples interpretaciones.
2. No usar términos **vagos**. Evitar el uso de términos vagos, ya que pueden generar confusión en los encuestados. En lugar de preguntar «¿le duele a menudo o con frecuencia?», es preferible especificar el intervalo de tiempo, como «¿le duele diaria o semanalmente?» para obtener respuestas más precisas.
3. Usar lenguaje claro y **sencillo**. Evitar términos técnicos o jerga especializada y utilizar palabras que sean comprensibles para la mayoría de los encuestados. Por ejemplo, en lugar de usar el término *menarquia*, se puede preguntar «¿a qué edad tuvo su primera regla?» para que sea más fácil de entender.
4. No formular preguntas en **negativo**, ya que pueden generar confusión y llevar a respuestas incorrectas. Es mejor plantearlas en afirmativo para que los encuestados puedan responder de manera clara y directa.
5. No evaluar **más de un concepto** o tema con una sola pregunta. Al evaluar múltiples conceptos en una sola pregunta, se corre el riesgo de obtener

respuestas confusas o incompletas. Es mejor dividir los conceptos en preguntas separadas para obtener respuestas más precisas y específicas.

6. Utilizar preguntas **cerradas** para cuestiones **personales**. Esto puede facilitar la respuesta de los encuestados y garantizar la coherencia en la recopilación de datos.
7. No olvidar **posibles alternativas** en las preguntas cerradas. Las opciones predefinidas en preguntas cerradas pueden no abarcar todas las posibilidades, en cuyo caso, se puede recurrir a preguntas semicerradas.
8. Utilizar alternativas mutuamente **excluyentes** y **ordenarlas** de manera lógica. Cuando las preguntas no sean mutuamente excluyentes hay que indicarlo expresamente en el enunciado. Además, se deben ordenar las opciones de manera lógica, ya sea en orden ascendente o descendente, para evitar sesgos en las respuestas.
9. No utilizar intervalos **solapados**. Al utilizar escalas de valoración o intervalos en las preguntas, se debe evitar que estos se solapen. Cada opción de respuesta debe ser claramente definida y no debe haber ambigüedad en cuanto a qué opción elegir.
10. Evitar preguntas **dirigidas**, que sugieran una respuesta específica o que orienten la respuesta del encuestado. Las preguntas deben ser imparciales y permitir que los encuestados expresen libremente su opinión o experiencia.
11. Identificar las **unidades** de medida. Si se hacen preguntas que requieren una respuesta numérica, es importante especificar las unidades de medida correspondientes. Esto ayuda a evitar confusiones y garantiza que los encuestados proporcionen respuestas coherentes y uniformes.

4.5. Orden de las preguntas

La introducción del cuestionario debe proporcionar una breve explicación sobre la naturaleza y el propósito del estudio. Esto ayudará a los participantes a comprender la importancia de su participación y a implicarse para el resto de la encuesta.

Durante toda la encuesta, se deben ordenar las preguntas de manera **lógica** y coherente. Esto facilita la comprensión por parte de los participantes y ayuda a mantener el **flujo** de la encuesta. En el caso de que haya preguntas abiertas y cerradas sobre el mismo tema, se recomienda plantear las **preguntas abiertas al inicio**. Esto evita que las alternativas de las preguntas cerradas influyan en las respuestas y permite obtener una mayor variedad de opiniones y perspectivas.

Sin embargo, se recomienda comenzar la encuesta con preguntas **sencillas** que sean fáciles de responder. Estas preguntas pueden servir como una forma de calentamiento para que los encuestados se sientan cómodos y se involucren en el proceso. Al iniciar la encuesta hay que evitar preguntas abiertas complicadas o cerradas con demasiadas opciones, ya que pueden resultar confusas o abrumadoras para los participantes. Es preferible utilizar preguntas **neutrales** y evitar aquellas que impliquen acuerdo o desacuerdo.

Es importante captar el **interés** de los encuestados desde el principio, por lo que se deben incluir preguntas que sean relevantes y atractivas para ellos. Esto ayudará a mantener su atención y motivación a lo largo de la encuesta.

Es recomendable incluir las preguntas sociodemográficas al final del cuestionario. También se pueden abordar preguntas delicadas, más **personales o conflictivas**, en esta etapa ya que podrían generar resistencia o desinterés por parte de los encuestados si se colocan al inicio.

4.6. Pautas generales para todo el cuestionario

1. Utilizar **recordatorios o instrucciones** específicas para el entrevistador, resaltándolos en cursiva. Estas instrucciones pueden incluir pautas sobre cómo abordar preguntas abiertas y cerradas del mismo tema, para asegurar que las alternativas de las preguntas cerradas no influyan en la respuesta del encuestado.
2. Especificar claramente **qué hacer ante una pregunta filtro**. Esto puede incluir indicaciones sobre cómo dirigir a los encuestados a un flujo de preguntas alternativo si no cumplen con los criterios establecidos. Estas instrucciones deben ser claras, precisas y estar resaltadas para evitar confusiones.
3. Asignar siempre el **mismo código** o etiqueta invariable para una misma opción al utilizar opciones de respuesta en preguntas cerradas. Esto ayuda a mantener la coherencia y facilita el análisis de los datos recopilados, ya que las respuestas se pueden categorizar de manera coherente.
4. Idealmente, **poner la opción de respuesta a la izquierda** en preguntas cerradas, especialmente cuando las preguntas son cortas. Esto facilita la lectura y la selección de las opciones para los encuestados. Sin embargo, en el caso de preguntas largas, colocar las opciones a la derecha puede ayudar a evitar errores al leer y seleccionar las respuestas.
5. Usar **títulos** o encabezados cuando se cambia de tema para evitar que el cuestionario se perciba como un listado interminable de preguntas y rom-

per la monotonía. Esto ayuda a que los encuestados se enfoquen en diferentes áreas temáticas y facilita la navegación a lo largo del cuestionario.

6. Utilizar **diferentes formatos de letra** para preguntas y respuestas, como negrita o cursiva, para hacer que las preguntas y respuestas se destaquen visualmente. Esto ayuda a identificar fácilmente las preguntas y las respuestas, lo que contribuye a una comprensión clara y una respuesta adecuada.
7. Agrupar la información por áreas temáticas. Esto facilita la navegación y comprensión del cuestionario para los encuestados, ya que pueden centrarse en un tema específico antes de pasar al siguiente. Además, ayuda al entrevistador a mantener un flujo lógico durante la administración del cuestionario.

5. Codificación

La codificación es un paso fundamental en el análisis de los datos de un cuestionario. Consiste en asignar categorías o valores numéricos a las respuestas obtenidas, lo cual nos permite realizar análisis posteriores. A continuación, se detallan aspectos importantes que considerar en este proceso de codificación.

5.1. Tipo de variable

Dependiendo del tipo de variable que estemos codificando, ya sea categórica (nominal u ordinal) o cuantitativa (discreta o continua), utilizaremos diferentes métodos y criterios.

En el caso de variables categóricas, se asignan categorías o etiquetas a las respuestas, con el objetivo de representar las diferentes opciones de respuesta. Estas categorías pueden ser nominales (no tienen un orden específico) u ordinales (tienen un orden jerárquico).

5.2. Escalas de valoración

Para medir conceptos subjetivos, los psicólogos han desarrollado diversas escalas de valoración que permiten asignar valores numéricos a las respuestas, como las ampliamente utilizadas de **Likert**, **Guttman**, **Thurstone** y **Osgood**. Estas escalas son herramientas utilizadas en los cuestionarios para facilitar la cuantificación de

conceptos abstractos y subjetivos y permitir analizarlos de manera más precisa y objetiva, pues estas escalas miden la intensidad o el grado de acuerdo o desacuerdo de una persona con respecto a un determinado enunciado o afirmación.

- Escala de **Likert**

La escala de Likert es una de las más utilizadas en la investigación social y psicológica. Consiste en una serie de afirmaciones o enunciados a los que los participantes deben responder indicando su nivel de acuerdo o desacuerdo en una escala de varios puntos, cinco o siete habitualmente. Por ejemplo, *Totalmente en desacuerdo*, *En desacuerdo*, *Neutral*, *De acuerdo* y *Totalmente de acuerdo*, aunque es común que solo se indiquen las alternativas extremas. Esta escala permite obtener una medida cuantitativa del grado de acuerdo o desacuerdo de una persona con respecto a una afirmación determinada. Es una herramienta ampliamente utilizada para cuantificar actitudes y conductas en estudios de investigación y consiste en un rango de 10 a 20 frases que expresan actitudes favorables o desfavorables hacia un tema específico. El número de frases con actitudes favorables y desfavorables suele ser similar, lo que permite obtener una medida equilibrada de las actitudes.

En la escala de Likert, se asigna una puntuación más alta a los ítems considerados favorables. Por ejemplo, si la escala consta de cinco puntos, la opción *Totalmente de acuerdo* puede recibir una puntuación de 5, mientras que la opción *Totalmente en desacuerdo* puede tener una puntuación de 1. Las opciones intermedias se puntúan de manera consecutiva.

Para obtener una puntuación total en la escala de Likert, se suman los puntos asignados a cada opinión o respuesta seleccionada por el participante. Por ejemplo, si una persona selecciona *De acuerdo* (puntuación de 4) en cinco frases favorables y *En desacuerdo* (puntuación de 2) en dos frases desfavorables, su puntuación total sería la suma de los puntos de cada opinión seleccionada, 24 puntos.

La puntuación total en la escala de Likert puede variar dependiendo del número de ítems en la escala y del rango de puntuaciones asignadas a cada opción de respuesta. No hay una puntuación *correcta* o *incorrecta* en esta escala, ya que su objetivo es medir el grado de acuerdo o desacuerdo del participante con respecto a las afirmaciones presentadas.

- Escala de **Guttman**

La escala de Guttman, también conocida como escala de acumulación, se utiliza para medir la actitud de una persona hacia una serie de enunciados relacionados con un tema específico. Los enunciados están cuidadosamente ordenados de manera que cada afirmación sea más intensa o extrema

que la anterior. Por ejemplo, si estamos midiendo la actitud hacia la comida rápida, los enunciados podrían comenzar con afirmaciones suaves como *A veces como comida rápida* y luego progresar hacia enunciados más extremos como *Prefiero la comida rápida a cualquier otra opción de comida*.

Los participantes deben indicar los enunciados que están más de acuerdo con su opinión. Se asume que, si una persona está de acuerdo con un enunciado extremo, también está de acuerdo con los enunciados menos extremos. Esto se basa en la premisa de que una persona que sostiene una opinión extrema también acepta todas las afirmaciones menos extremas que se derivan de ella.

La escala de Guttman se utiliza para identificar un patrón de respuestas unidimensional, es decir, para determinar si los participantes siguen una secuencia lógica y coherente de respuestas que van desde las afirmaciones menos extremas hasta las más extremas. Esto puede ayudar a clasificar a los participantes en diferentes niveles de una actitud o creencia, lo que permite obtener una comprensión más detallada y precisa de las opiniones de las personas sobre un tema en particular.

A pesar de que la escala de Guttman ha sido ampliamente utilizada en el pasado, su uso ha disminuido debido a la aparición de técnicas de medición más sofisticadas y precisas. Sin embargo, su enfoque en la secuencia ordenada de respuestas sigue siendo valioso para comprender la estructura y dirección de las actitudes y creencias de las personas.

- Escala de **Thurstone**

La escala de Thurstone, desarrollada por el psicólogo Louis L. Thurstone, es una técnica de medición que usa la teoría de los juicios comparativos y se basa en la premisa de que los participantes pueden realizar juicios comparativos más fácilmente que asignar valores absolutos a los enunciados. Esta escala se utiliza para medir la actitud de los participantes hacia una serie de enunciados o ítems. En ella, los participantes se enfrentan a una serie de pares de enunciados cuidadosamente seleccionados para cubrir todo el rango de opiniones posibles en la dimensión que se está evaluando. Cada par de enunciados representan puntos opuestos en una dimensión específica. Por ejemplo, si estamos evaluando la actitud hacia los teléfonos móviles, un par de enunciados podría ser *Los teléfonos móviles son esenciales para la vida moderna* y *Los teléfonos móviles son una amenaza para la privacidad*.

Los participantes deben comparar los pares de enunciados y seleccionar cuál representa mejor su opinión o posición. Pueden usar escalas de valoración numérica, de 1 a 7 por ejemplo, para indicar su preferencia. Con base

en las respuestas de los participantes, se pueden calcular puntuaciones para cada enunciado y determinar el orden o nivel de acuerdo de los enunciados en la dimensión evaluada.

El cálculo de las puntuaciones se realiza utilizando métodos estadísticos, como la media ponderada, que tienen en cuenta tanto la posición del enunciado en relación con otros como la coherencia de las respuestas de los participantes. De esta manera, se obtiene una medida cuantitativa del acuerdo o desacuerdo de los participantes con cada enunciado y se puede establecer un orden o clasificación de los enunciados en función de la fuerza de acuerdo o desacuerdo de los participantes. Esto permite identificar patrones y tendencias en las actitudes y opiniones de las personas sobre el tema evaluado.

- Escala de **Osgood**

La escala de Osgood, también conocida como escala semántica diferencial, se utiliza para medir las actitudes o evaluaciones de los participantes hacia un objeto o concepto específico. Los participantes deben indicar su grado de acuerdo o desacuerdo en una serie de atributos o características opuestas, utilizando una escala de varios puntos. Por ejemplo, se pueden utilizar pares de adjetivos opuestos como *bueno* y *malo*, *agradable* y *desagradable*. Esta escala permite capturar la evaluación subjetiva de un objeto o concepto en múltiples dimensiones y proporciona información detallada sobre las percepciones de los participantes.

La elección de la escala adecuada depende del objetivo de la pregunta, el tipo de constructo que se desea medir y las características de los participantes.

En la tabla se comparan las cuatro escalas descritas.

Tabla 1. Comparación de las escalas de valoración

Escala	Método de medición	Base teórica	Ejemplo																												
Likert	Respuestas en una escala de puntos.	Teoría de actitudes y nivel de acuerdo.	Estoy satisfecho/a con la calidad de atención recibida por el personal médico en el hospital. a) Totalmente en desacuerdo. b) En desacuerdo. c) Neutral. d) De acuerdo. e) Totalmente de acuerdo.																												
Guttman	Respuestas ordenadas y acumulativas (enunciados ordenados por intensidad de acuerdo).	Teoría de actitudes y secuencia lógica de respuestas.	Nivel de adherencia al tratamiento para pacientes con diabetes. 1. A veces olvido tomar mi medicación para la diabetes. 2. Generalmente tomo mi medicación para la diabetes según lo indicado. 3. Siempre tomo mi medicación para la diabetes según lo indicado y mantengo mis citas médicas. 4. Además de tomar mi medicación y asistir a mis citas, también sigo una dieta y rutina de ejercicio recomendadas por mi médico.																												
Thurstone	Respuestas en comparación de pares (solo puede seleccionar en función de su acuerdo o desacuerdo con la afirmación).	Teoría de los juicios comparativos y preferencia.	Actitud hacia la vacunación contra una enfermedad infecciosa. a) La vacunación es una medida efectiva para prevenir la propagación de la enfermedad. <table><tr><td>1 (nada de acuerdo)</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7 (totalmente de acuerdo)</td></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table> b) La vacunación no es necesaria y puede tener efectos secundarios negativos. <table><tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr><tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table>	1 (nada de acuerdo)	2	3	4	5	6	7 (totalmente de acuerdo)								1	2	3	4	5	6	7							
1 (nada de acuerdo)	2	3	4	5	6	7 (totalmente de acuerdo)																									
1	2	3	4	5	6	7																									
Osgood	Respuestas en una escala de puntos entre dos atributos opuestos (extremos).	Teoría de las evaluaciones semánticas y atributos opuestos.	Percepción del bienestar emocional en pacientes con trastornos mentales. <table><tr><td></td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td></td></tr><tr><td>Muy malo</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Muy bueno</td></tr><tr><td>Muy inestable</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Muy estable</td></tr></table> Los pacientes marcarían un punto en la escala para expresar su percepción subjetiva sobre su bienestar emocional en el contexto de su trastorno mental.		1	2	3	4	5	6	7		Muy malo								Muy bueno	Muy inestable								Muy estable	
	1	2	3	4	5	6	7																								
Muy malo								Muy bueno																							
Muy inestable								Muy estable																							

5.3. La escala visual analógica (EVA)

Es una herramienta de medición utilizada en la clínica y en los estudios de investigación para cuantificar las percepciones subjetivas de los individuos sobre un determinado constructo que se basa en la representación visual de una línea continua, generalmente horizontal, que tiene dos extremos claramente definidos.

En la escala visual analógica, a los participantes se les pide que marquen un punto en la línea continua para indicar su posición o grado de intensidad en relación con el constructo que se está evaluando de acuerdo con su percepción o experiencia personal. La ubicación del punto marcado por el participante se convierte en una medida cuantitativa que puede ser utilizada para comparar y analizar las respuestas de diferentes individuos o para medir cambios a lo largo del tiempo en una misma persona.

Las escalas visuales analógicas pueden variar en longitud, pero generalmente se presentan en una escala de 10 centímetros. El extremo izquierdo de la línea suele representar el extremo negativo o la ausencia del constructo evaluado, mientras que el extremo derecho representa el extremo positivo o la máxima intensidad del constructo.

Una ventaja clave de la escala visual analógica es su facilidad de uso y comprensión. La representación visual de la escala permite a los participantes comprender rápidamente el concepto y proporcionar una respuesta basada en su propia percepción. Además, tiene la capacidad de capturar tanto la intensidad como la dirección de una experiencia o constructo, lo que la hace versátil y aplicable a una amplia gama de contextos.

En el ámbito clínico, se ha utilizado para evaluar y cuantificar el dolor, la ansiedad, la depresión, la calidad de vida relacionada con la salud y otros constructos subjetivos. También se ha utilizado en investigaciones para medir la satisfacción del paciente, la percepción de la eficacia de un tratamiento o la evaluación de experiencias emocionales.

La interpretación de los resultados de una escala visual analógica puede variar según el contexto y el constructo evaluado. Por lo tanto, se deben considerar los límites claramente definidos en la escala para garantizar una interpretación adecuada y uniforme de las respuestas.

6. Administración de cuestionarios

El objetivo principal al administrar un cuestionario es recoger los datos del mayor porcentaje posible de participantes. Para lograrlo, se recomienda:

1. El desarrollo del **manual** del encuestador que incluya los objetivos del estudio y la técnica de recopilación de información. Este documento servirá como guía para el encuestador, con instrucciones claras sobre cómo llevar a cabo la administración del cuestionario.
2. **Formación** y capacitación del encuestador mediante sesiones teórico-prácticas en las que se le enseñará cómo abordar a los participantes, realizar las preguntas de manera adecuada y registrar las respuestas de manera precisa.

6.1. Método de administración

Existen diferentes métodos de administración para los cuestionarios, cada uno con sus ventajas e inconvenientes. Dos tipos comunes de administración son la autoadministración y la administración a través de un entrevistador o entrevistadora.

La **autoadministración** es un método en el que los participantes completan el cuestionario por sí mismos, lo que ofrece ciertas ventajas. En primer lugar, permite a los participantes completar el cuestionario a su propio ritmo y en el momento que prefieran, lo que puede resultar más conveniente para aquellos que prefieren la intimidad al responder. Además, los participantes tienen la posibilidad de reflexionar sobre las preguntas y proporcionar respuestas más cuidadosas y sinceras. Sin embargo, también tiene sus inconvenientes. Existe la posibilidad de que los participantes no completen o no devuelvan el cuestionario, lo que puede afectar la tasa de respuesta, con sus sesgos consiguientes. Además, al no haber un entrevistador presente, los participantes no tienen la oportunidad de obtener aclaraciones adicionales o de hacer preguntas sobre las instrucciones o los elementos del cuestionario que puedan resultar confusos.

La administración a través de un **entrevistador** implica la presencia de una persona que guía y administra el cuestionario a los participantes. Este método tiene sus ventajas. El entrevistador puede brindar claridad sobre las preguntas o instrucciones, lo que ayuda a minimizar las confusiones y asegura que los participantes comprendan adecuadamente el propósito del cuestionario. Además, la presencia del entrevistador puede motivar a los participantes a completar el cuestionario y garantizar que todas las preguntas sean respondidas. Sin embargo, también tiene sus inconvenientes. La presencia de un entrevistador puede influir en las respuestas de los participantes debido al sesgo de deseabilidad social: los participantes pueden sentir la presión de dar respuestas que consideren más aceptables o socialmente

deseables en lugar de responder de manera sincera. Además, este método requiere asignar recursos adicionales para contratar y capacitar a los entrevistadores, lo que puede suponer un coste adicional y un tiempo de preparación más prolongado.

La elección entre la autoadministración y la administración a través de un entrevistador dependerá de varios factores, como la naturaleza del estudio, los recursos disponibles y las preferencias de los investigadores.

6.2. La importancia del entrevistador

El entrevistador desempeña un papel crucial en la administración del cuestionario. Su habilidad para establecer un buen vínculo con los participantes y seguir las instrucciones correctamente influye en la calidad de los datos recopilados. Además, el entrevistador debe valorar la entrevista y registrar su hora de finalización.

6.3. Entrevista asistida por ordenador

La entrevista asistida por ordenador es un método utilizado para administrar cuestionarios electrónicos, que mejora la calidad de los datos recopilados respecto a la administración en papel impreso. Hay varios métodos de entrevista asistida por ordenador, como CAPI, CATI, CASI y CAWI.

- La **CAPI** (entrevista personal asistida por ordenador) es una técnica en la cual el **entrevistador** utiliza un dispositivo electrónico, como un ordenador portátil, una tableta o un teléfono inteligente, para realizar las preguntas al encuestado. Se pueden incluir pantallas de ayuda y mensajes de error para facilitar la comprensión.
- La **CASI** (**autoentrevista** asistida por ordenador) es similar a la CAPI, pero sin la presencia de un entrevistador. Hay dos tipos: vídeo-CASI y audio-CASI. Estos métodos pueden facilitar la respuesta a preguntas delicadas dada la ausencia de un entrevistador. Sin embargo, puede introducir un sesgo de selección, ya que solo aquellos con conocimientos informáticos y acceso a Internet podrán participar.

La **vídeo-CASI** implica que los encuestados lean las preguntas en la pantalla y respondan utilizando un teclado u otro dispositivo de entrada. Este método puede adaptar las preguntas en función de las respuestas anteriores, avisar de las respuestas fuera de rango y estandarizar la entrevista para disminuir los errores. Sin embargo, puede requerir habilidades de lectura y puede imponer una carga visual y de lectura adicional al formato de papel impreso.

La **audio-CASI** formula las preguntas de manera audible para los encuestados y les permite escuchar y responder mientras las preguntas están visibles

en la pantalla. Este método supera las barreras de alfabetización y ofrece una intimidad similar a la de los cuestionarios en papel autoadministrados.

- La **CAWI** (entrevista **web** asistida por ordenador) se basa en un cuestionario en línea en el que los encuestados responden a las preguntas alojadas en un sitio web. Permite la inclusión de elementos visuales y auditivos en el cuestionario para aumentar el interés y la participación del entrevistado. Es más barata de administrar y permite recopilar datos en una base de datos central para un análisis más rápido. Sin embargo, la tasa de respuesta puede ser limitada debido a la falta de acceso a Internet y a que algunas personas pueden ser menos receptivas a los cuestionarios en línea.
- La **CATI** (entrevista **telefónica automatizada** por ordenador) es una técnica en la que un ordenador con capacidad de reconocimiento de voz formula preguntas a los encuestados y almacena sus respuestas. Sigue un guion cuya lógica puede adaptarse según las respuestas y la información conocida sobre el participante.

Las entrevistas asistidas por ordenador ofrecen ventajas en términos de calidad de datos, flexibilidad en la presentación de preguntas, reducción de costes y facilidad de corrección de errores. Sin embargo, también presentan desafíos relacionados con el sesgo de selección y la falta de acceso o receptividad de ciertos grupos de población.

6.4. Peculiaridades de las encuestas telefónicas

Las encuestas telefónicas presentan ciertas peculiaridades que las distinguen de otros métodos de recopilación de datos. En comparación con las encuestas personales, suelen tener un menor coste y son más rápidas de realizar. Además, suelen tener un mayor porcentaje de respuesta en comparación con las encuestas autocumplimentadas. Sin embargo, también presentan dificultades. Son más complejas de llevar a cabo, ya que algunos factores, como el tono de voz, pueden inducir a sesgos en las respuestas de los participantes.

Para aumentar la tasa de respuesta en las encuestas telefónicas, se recomienda utilizar preguntas breves y sencillas, así como un cuestionario de duración máxima de 10 minutos. También es importante evitar preguntas íntimas que puedan generar incomodidad en los participantes. Además, se considera importante enviar una carta al encuestado antes de realizar la encuesta telefónica para agradecer su colaboración, explicar los objetivos del estudio, mencionar la institución que lo realiza y proporcionar información sobre el uso que se hará de los datos recopilados. Esta comunicación previa contribuye a establecer una relación de confianza con el participante y aumenta la probabilidad de que acceda a participar en la encuesta telefónica.

7. Conclusiones

Los cuestionarios epidemiológicos son herramientas fundamentales en la investigación epidemiológica, que se utilizan para recopilar datos sobre la presencia, distribución y factores de riesgo de enfermedades o el estado de salud de una población. Su objetivo principal es obtener información precisa y veraz que permita identificar patrones de enfermedad, comprender sus causas y determinantes, evaluar la carga de la enfermedad en la población y diseñar intervenciones y políticas de salud pública efectivas.

Estos cuestionarios tienen diversos usos en el campo de la salud, como la investigación de brotes de enfermedades, los estudios de prevalencia, la investigación de factores de riesgo, la evaluación de programas de salud y la vigilancia de enfermedades.

Al seleccionar un cuestionario, es importante conocer los cuestionarios existentes y evaluar su validez, fiabilidad y utilidad. Se recomienda hacer uso de cuestionarios validados y revisar las guías de los cuestionarios existentes para seleccionar el más adecuado. Además, la adaptación cultural de un cuestionario es esencial para garantizar su comprensión y adecuación a la población objetivo.

La validación de un cuestionario implica evaluar su viabilidad, validez y fiabilidad, para lo que se utilizan diferentes métodos. Además, el cuestionario debe estar adaptado culturalmente a la población de destino en contenido, formato e idioma.

El diseño del cuestionario debe considerar el tipo de información a recopilar, la estructura, la función, la redacción y el orden de las preguntas. Es importante seguir pautas generales y utilizar cuestionarios con propiedades de medida demostradas para garantizar la calidad de los datos recopilados.

Una vez recopiladas las respuestas, la codificación de los datos es un paso esencial en el análisis del cuestionario. La asignación de categorías o valores numéricos a las respuestas permite realizar análisis posteriores y obtener una visión general y comparativa de las respuestas de los participantes.

Agradecimientos

A Esther García García-Esquinas y Beatriz Pérez Gómez por su generosidad genuina al compartir sus materiales y conocimientos y, por tanto, enriquecer la perspectiva de este capítulo.

8. Referencias

1. Badia X, Salamero M, Alonso J. La medida de la salud: guía de escalas de medición en español. 4.^a edición, Barcelona: Tecnología y Ediciones del Conocimiento. 2007.
2. De Yébenes Prous MJ, Rodríguez Salvanés F, Carmona Ortells L. Sensibilidad al cambio de las medidas de desenlace. *Reumatol Clin*. 2008;4:240-7.
3. Edwards P. Questionnaires in clinical trials: guidelines for optimal design and administration. *Trials*. 2010;11:2.
4. Jeanrie C, Bertrand R. Translating tests with the International Test Commission's Guidelines: keeping validity in mind. *Eur J Psychol Assess*. 1999;15:277-83.
5. Fink A. Epidemiological field work in population-based studies. En Ahrens W y Pigeot I (Eds). *Handbook of epidemiology*. 2nd edition. New York: Springer. 2014:577-611.
6. Laher S, editor. *International histories of psychological assessment*. Cambridge: Cambridge University Press. 2022.
7. Lauffer A, Solé L, Bernstein S, Lopes MH, Francisconi CF. Cómo minimizar errores al realizar la adaptación transcultural y la validación de los cuestionarios sobre calidad de vida: aspectos prácticos. *Rev Gastroenterol Mex*. 2013;78:159-76.
8. Lira MT, Caballero E. Adaptación transcultural de instrumentos de evaluación en salud: historia y reflexiones del por qué, cómo y cuándo. *Rev Med Clin Las Condes*. 2020;31:85-94.
9. Martín Arribas C. Diseño y validación de cuestionarios. *Matronas Profesión*. 2004;5:17.
10. McDowell I. *Measuring health: a guide to rating scales and questionnaires*. 3rd ed. Oxford: Oxford University Press. 2006.
11. Nieuwenhuijsen MJ. Design of exposure questionnaires for epidemiological studies. *Occupational and Environmental Medicine*. 2005;62:272-80.
12. Ortiz-Gutiérrez S, Cruz-Avelar A. Proceso de traducción y adaptación cultural de instrumentos de medición en salud. *Actas Dermosifiliogr*. 2018;109:202-6.
13. Ramada-Rodilla JM, Serra-Pujadas C, Delclós-Clanchet GL. Adaptación cultural y validación de cuestionarios de salud: revisión y recomendaciones metodológicas. *Salud pública de México*. 2013;55:57-66.
14. Streiner DL, Norman GR, Cairney J. *Health measurement scales. A practical guide to their development and use*. 5th edition. Oxford: Oxford University Press. 2015.

Capítulo 17.

Inferencia causal en epidemiología

JOSÉ RAMÓN BANEGAS BANEGAS, MIGUEL ÁNGEL ROYO BORDONADA
y FERNANDO RODRÍGUEZ ARTALEJO

1. Introducción

2. Filosofía de la inferencia causal

3. Modelos de causalidad en epidemiología

3.1. Modelo determinista modificado

3.2. Modelo probabilístico

3.3. Método contrafáctico

3.4. Causación a múltiples niveles: Modelo de red

4. Criterios de causalidad

4.1. Criterios de Hill

4.2. Otros criterios de causalidad

5. Práctica de la inferencia causal

6. Referencias

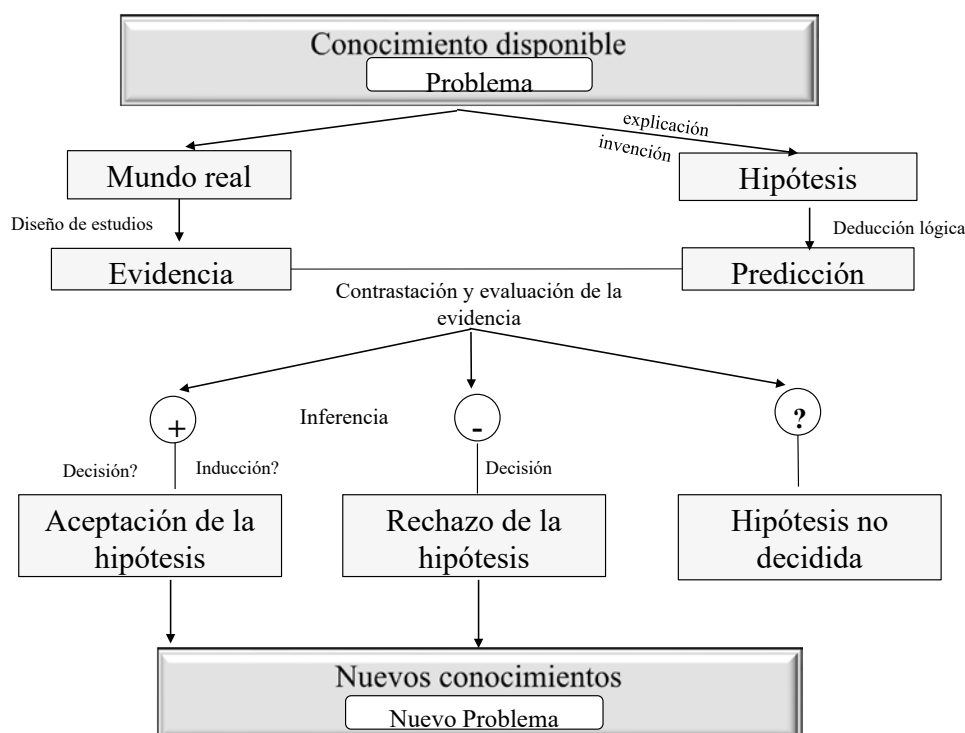
ANEXO I. Centro de Oxford para la medicina basada en pruebas: niveles de prueba

1. Introducción

La epidemiología es el estudio de la distribución y los determinantes de los estados o acontecimientos relacionados con la salud en poblaciones específicas y la aplicación de este estudio al control de problemas de salud. (1) Por tanto, la inferencia causal es uno de los cometidos esenciales de la epidemiología. Para MacMahon, el propósito más importante de la epidemiología es adquirir conocimiento de las causas de las enfermedades que no son prevenibles actualmente. (2)

Cualquier persona adulta dispone de conocimientos básicos sobre causalidad, que ha adquirido desde la primera infancia a base de ensayo y error. (3) Los niños experimentan con el mundo que les rodea como si fueran científicos: a partir de las percepciones que reciben del exterior, imaginan posibles explicaciones (“hipótesis”) sobre el funcionamiento del mundo y las relaciones entre los acontecimientos percibidos, que después verifican mediante ensayos repetidos, como cuando comprueban una y otra vez, hasta abandonar toda sombra de duda, que al accionar los interruptores situados a la entrada de las habitaciones se hace la luz, o que acercar la mano a cierta distancia del fuego durante cierto tiempo es doloroso. Este conocimiento naturalmente adquirido nos permite comprender la diferencia entre asociación y causalidad; de otra forma, nadie en su sano juicio aceptaría ingresar en un hospital, ya que la probabilidad de morir es mucho mayor entre las personas hospitalizadas que en la población no hospitalizada. (4)

Los conocimientos sobre causalidad naturalmente adquiridos son demasiado rudimentarios como para sentar las bases de una teoría científica y un método sistemático para la elaboración y comprobación de hipótesis o explicaciones causales. La búsqueda de agentes infecciosos por los primeros cazadores de microbios, con la pretensión de identificarlos como causas de enfermedades, episodio que podemos transferir a los actuales cazadores de factores de riesgo o de genes, (5) resulta de poca utilidad fuera de un contexto metodológico y sustantivo sobre causalidad. (6) La inferencia causal se comprende mejor si la ubicamos en el marco general, más amplio y adecuado de la explicación causal (figura 1). Explicación, comprobación e inferencia siempre entran en juego en el método científico, aunque dependiendo de cuál es nuestra situación problemática: la búsqueda de una hipótesis (causal) explicativa, la comprobación de una hipótesis ya formulada o la inferencia sobre una hipótesis ya contrastada.

Figura 1. Método de investigación científica.

Fuente: Modificado de Popper, Bunge y Giere.

Así pues, establecer las causas de los fenómenos de interés es uno de los principales objetivos —y problemas— de la actividad científica. Este capítulo tiene por objeto responder a dos preguntas: ¿qué es una causa?, para lo cual revisamos los modelos o representaciones formales de causalidad de mayor interés en epidemiología y salud pública, y ¿cómo reconocemos en la práctica que una variable o fenómeno concreto es la causa de otro?

2. Filosofía de la inferencia causal

El tratamiento filosófico de la causalidad se remonta a la Antigüedad. Ya Hipócrates (siglo V a. de C.) sugirió que el desarrollo de la enfermedad humana podría estar relacionado con factores ambientales; es decir, enunció que las enfermedades tenían causas. Aristóteles planteó la necesidad de un procedimiento inductivo para acceder a las leyes universales que rigen la naturaleza a partir de la observación repetida de los fenómenos naturales, la inducción por enumeración. Pero no fue hasta el siglo XVII (1620) cuando Bacon, en su *Novum Organum*, de-

sarrolló un método científico del razonamiento inductivo, como base más segura para acceder al conocimiento, frente al razonamiento deductivo predominante hasta entonces, basado en ideas preconcebidas y dogmas, cuya validez depende del contenido de las premisas, de modo que la conclusión, implícita en las premisas, solo será cierta si estas también lo son. (7) Bacon propone un ascenso gradual a partir de la observación y descripción de los hechos, su tabulación y clasificación, mediante la eliminación de todo aquello que no está conectado con el fenómeno bajo estudio, dando por cierto lo que sobrevive a este cribado. Y llama la atención sobre el mayor peso de un resultado negativo: mientras que la observación de una nueva especie de mamífero que resulta ser vivíparo contribuye débilmente a confirmar la tesis “los mamíferos son vivíparos”; el descubrimiento del equidna, un mamífero ovíparo, basta por sí solo para demostrar la falsedad de la misma. Un siglo después, David Hume señaló que el razonamiento inductivo se basa en asumir que la naturaleza es uniforme (en el espacio y el tiempo), algo injustificable de forma lógica por el hecho de que hasta ahora haya sido así, lo que nos lleva al problema de la inducción: por muchas veces que veamos repetirse una secuencia de eventos, nunca puede excluirse la posibilidad de mera coincidencia. Nada puede asegurarnos que esa secuencia de eventos seguirá en el futuro el mismo patrón que ha seguido hasta ahora. (8) Es lo que Bertrand Russell denominó la falacia del consecuente: si el canto del gallo fuese la causa de que saliera el sol, entonces el sol debería salir después de que cante el gallo; como el sol sale efectivamente después de que el gallo cante, entonces el canto del gallo es la causa de la salida del sol. La solución de Karl Popper al problema de la inducción consiste en proponer un método científico que, en lugar de intentar verificar una hipótesis, lo que busca es probar la falsedad de la misma (refutaciónismo), retomando en cierto modo a Bacon, que otorgaba mayor peso a los resultados negativos. Se trata de un método hipotético-deductivo en el que, partiendo de la hipótesis de una relación causal (por ejemplo, el tabaco causa cáncer) y las condiciones iniciales (espacio-temporales) de un estudio no sesgado en una población determinada, se establece una predicción mediante la aplicación de la lógica deductiva. (9) Si los resultados del estudio son incompatibles con la predicción, se rechaza la hipótesis (un día el gallo muere y, no obstante, el sol sale al día siguiente); en caso contrario, la hipótesis queda provisionalmente corroborada, que no verificada, ya que los hechos, aun siendo compatibles con nuestra hipótesis, podrían ser debidos a hipótesis alternativas (el sol sale cada mañana porque la Tierra gira en torno a su eje y no porque el gallo cante), y sujeta a nuevos test de falsificación. Si la hipótesis no es determinista, sino probabilística (el tabaco aumenta la probabilidad de desarrollar cáncer), el razonamiento es el mismo, pero la hipótesis ya no es estrictamente falsable, sino en términos probabilísticos, pues nunca podemos descartar que, aun siendo cierta la hipótesis, la muestra seleccionada presente, por puro azar, una desviación

importante respecto a las predicciones. En la práctica, esto se resuelve mediante reglas de decisión, como el valor p (nivel de error que consideramos aceptable) y el grado de concordancia entre estudios, pues al aumentar el número de observaciones incompatibles con la hipótesis, la probabilidad de que estas se deban al azar se hace más despreciable. Cuando los resultados son compatibles con la hipótesis, nada podemos afirmar de forma lógica acerca de la misma, ya que el valor p es la probabilidad de los datos, no de la hipótesis. Sin embargo, el objetivo último de la ciencia es la búsqueda de la verdad objetiva, de leyes universales, válidas en toda circunstancia, lugar y tiempo. Por tanto, la ciencia sigue una dirección inductiva, de los datos a la hipótesis, aunque para ello se apoye en una lógica deductiva, que proporciona apoyo probabilístico únicamente a la parte del contenido de la hipótesis incluida en los propios datos. Este proceso inductivo tiene lugar en las fases iniciales del ciclo de la investigación, con la generación de hipótesis, inventadas por el investigador a partir de observaciones previas, y en las finales, con la evaluación de la hipótesis. No podemos inducir la probabilidad de la hipótesis, pero podemos juzgar la misma en base a argumentos críticos y reglas metodológicas, mediante pruebas contrastables, reproducibles y discutibles intersubjetivamente. Por ejemplo, el argumento de la concordancia: la hipótesis será más verosímil cuando ha sido corroborada en múltiples observaciones, realizadas en poblaciones, circunstancias, lugares y momentos diferentes, con varios diseños de estudios, pues ha sido sometida a exámenes duros que aumentan las posibilidades de refutarla. Otro argumento crítico es el de la verosimilitud biológica, ya que una teoría que especifica el mecanismo de acción es más rica en su explicación de la realidad y proporciona más oportunidades de contrastación y refutación de hipótesis vinculadas a la misma. Estos y otros argumentos críticos, tradicionalmente denominados criterios, que veremos con detalle más adelante, nos ayudan a valorar el grado en que los datos de los estudios realizados apoyan o no, y en qué medida, la hipótesis causal, y decidir, en la práctica, si la aceptamos o rechazamos —siempre de forma provisional—, ya que la acción es necesariamente dicotómica. La decisión se toma por un acuerdo o convención razonado, reglamentado y objetivo. En caso de duda, hay que renunciar a tomar un acuerdo y proseguir la comprobación y discusión. Precisamente, que sean decisiones y no conclusiones (lógicas) lo que en último término decide el destino de una hipótesis (causal) habla de que la evaluación de la hipótesis es tentativa, provisional, no concluyente o definitiva. (8,10)

El bayesianismo es una alternativa al refutacionismo donde, a diferencia de la estadística frecuentista del método de Popper, que calcula la probabilidad de los resultados observados dada la hipótesis, se calcula la probabilidad de la hipótesis dados los resultados observados (probabilidad inductiva), mediante una estadística basada en el teorema de Bayes, derivado por lógica deductiva de las leyes de probabilidad. Ya que no es posible probar de forma concluyente que una hipótesis

es cierta, aunque lo sea, la inferencia inductiva (bayesiana) permite cuantificar el grado de certeza de la misma proporcionado por los datos observados. La principal limitación del bayesianismo es que la validez de sus conclusiones depende de la certeza de sus premisas. Cuando se trata de hipótesis sobre las leyes naturales, es precisamente la veracidad o falsedad de las premisas lo que pretendemos demostrar, lo que nos lleva a un círculo vicioso. El bayesianismo aborda este problema mediante la asignación de un grado de certeza a las premisas. Si puedes asignar una probabilidad previa a la hipótesis (grado de certeza), puedes utilizar la inferencia bayesiana para derivar de forma lógica su probabilidad *a posteriori*. Como los investigadores pueden discrepar en la asignación de la probabilidad *a priori*, la inferencia Bayesiana ha sido tachada de subjetiva. Sin embargo, la distribución de probabilidad posterior es bastante insensible a variaciones en la distribución de la probabilidad previa y esta insensibilidad aumenta con el tamaño de la muestra, por lo que dos investigadores que parten de creencias en probabilidades previas muy diferentes pueden terminar convergiendo en la misma probabilidad posterior con la progresiva acumulación de datos). (11)

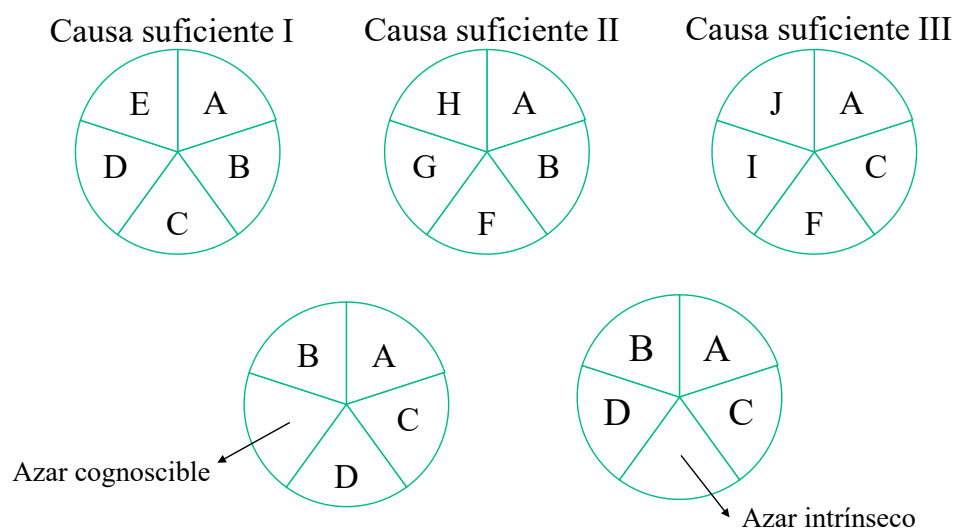
3. Modelos de causalidad en epidemiología

El término causa no es unívoco en el campo de la epidemiología. De forma concisa, una causa es un suceso, condición o característica que produce un efecto. (12) Pero esta definición es vaga y no se acomoda bien con la complejidad del fenómeno de la causalidad ni facilita la identificación de las causas particulares de los fenómenos de interés relacionados con la salud. A tales efectos, se han desarrollado dos enfoques del concepto de causación, el determinista y el probabilístico. De ahí que se utilicen los términos determinante y factor de riesgo para aludir al mismo concepto de causa. El concepto determinista de causa distingue entre causa necesaria, aquella condición sin la cual el efecto no puede producirse, y causa suficiente, aquella condición cuya presencia produce inexorablemente el efecto; mientras que el concepto probabilístico define una causa como aquella condición que aumenta la probabilidad de que se produzca un efecto. Con la ayuda del método contrafáctico, que permite distinguir entre causación y mera correlación, ambos modelos son útiles para identificar causas concretas y cuantificar la magnitud de su efecto en determinadas circunstancias, pero son incapaces de abarcar y desenmarañar toda la complejidad del fenómeno de la causalidad, para lo que se requiere un abordaje a múltiples niveles, propio del modelo de red.

3.1. Modelo determinista modificado

Algunos autores, como Rothman, propugnan un modelo general de causalidad que permite relacionar algunos de los principales principios epidemiológicos: fuerza de los efectos, interacción entre causas, proporciones de la enfermedad debidas a causas específicas, período de inducción, etc. El modelo propuesto es un modelo teórico —es decir, imaginario— pero que ayuda a visualizar cómo podrían ser las cosas y, por lo tanto, alienta a comprobarlas. (13) En este modelo, causa se define como la condición o característica que inicia o permite, sola o en conjunto con otras causas, una secuencia de sucesos que resulta en un efecto. Se trata de un modelo determinista modificado. Es determinista porque supone que una enfermedad puede tener una o más causas suficientes (CS), cada una de las cuales, por sí mismas, determina inexorablemente la aparición de la enfermedad en un individuo y una proporción determinada de casos de la enfermedad en la población. Cada CS (o mecanismo causal teórico) está compuesta de una constelación mínima de causas componentes (CC) que interactúan para producir un efecto (figura 2). Mínima significa que cada una de las CC es imprescindible para producir la enfermedad debida a una determinada CS. Por último, se denomina causa necesaria a aquella CC que está presente en todas las CS de una enfermedad, y sin la cual, por lo tanto, no se produce ningún caso de enfermedad.

Figura 2. Modelo causal determinista modificado.



Fuente: Modificado de Rothman, 1998.

Un problema de la concepción determinista es que, en el estado actual del conocimiento, para la mayoría (si no todas) de las enfermedades conocemos solo algunos de los componentes de las CS. Por eso, el modelo comentado atempera

el determinismo modificándolo, dando cabida al azar en forma de factores o CC todavía no identificados. No conocemos ninguna CS, ni siquiera del cáncer de pulmón. Así pues, que un individuo fumador, o expuesto a radón o asbestos, desarrolle o no cáncer de pulmón se debe a que actúen o no otros factores causales aún desconocidos, necesarios para completar una CS. Para una causa componente, en un contexto de causa suficiente incompleta —es decir, en presencia de azar—, solo puede cuantificarse su efecto en forma probabilística.

Apoyándonos en estos conceptos y en la figura 2, es posible visualizar cómo el modelo determinista modificado por el probabilismo acomoda algunos principios epidemiológicos. Primero, no es necesario conocer todas las CC de una CS para prevenir la enfermedad. Eliminando uno solo de los elementos de una CS se previenen todos los casos de enfermedad (o fracción etiológica [FE]) que esa CS origina. Así, si una CS de una enfermedad tiene cinco CC (a saber, A, B, C, D, E), y la FE de la CS es 20 %, la FE de “D” es al menos del 20 %, pues sin “D” la CS no se completa y, en consecuencia, la enfermedad no se produce. Eliminando “D” se evitarían al menos el 20 % de los casos de esa enfermedad, los causados por esa CS, incluso en presencia del resto de las CC (A, B, C y E), una cifra que podría ser superior en la medida en que “D” también fuese una CC de otras CS. Por ejemplo, la eliminación del uso de jeringuillas contaminadas (solo una causa componente) prevendría la aparición de un gran número de casos de sida, hepatitis viral, etc. Segundo, la suma de las FE de todas las CS de una enfermedad es siempre el 100 %. Sin embargo, la FE correspondiente a cada CC equivale a la suma de las FE de la(s) CS de las que forma parte. De este modo, si una enfermedad tiene tres CS, siendo la FE de la CS primera FE (I) = 20 %, y la de la tercera FE (III) = 50 %, entonces la FE (II) = 30 %. Y si la CS (I) la forman las CC A, B, C, D, E; la II está formada por A, B, F, G, H, y la III lo está por A, C, F, I, J, entonces:

$$FE(D) = FE(E) = FE(I) = 20 \%$$

$$FE(B) = FE(I) + FE(II) = 50 \%$$

$$FE(A) = FE(I) + FE(II) + FE(III) = 100 \%$$

$$FE(A) + FE(B) + FE(C) + \dots + FE(J) = 500 \%$$

Se colige que en salud pública deberían priorizarse aquellas CC que por estar implicadas en varias CS responderían de una mayor proporción de la enfermedad. Que la suma de las FE de todas las CC pueda superar el 100 % no significa obviamente que pueda prevenirse más del 100 % de una enfermedad. No hay límite superior a la suma de las FE. Solo la FE atribuible a una causa aislada no puede exceder el 100 %. Y solo una causa necesaria contribuirá al 100 %. Por consiguiente, para cualquier enfermedad, la suma de las FE de las CC es al menos del 100 %. Así, se ha declarado que los porcentajes de casos de cáncer atribuibles a factores

de riesgo son los siguientes: el 10-70 % a la alimentación insalubre, el 25-35 % al tabaco, el 10-40 % a los tóxicos laborales, el 5-10 % a los tóxicos ambientales y el 10-20 % al alcohol. Por su parte, el porcentaje de cardiopatías atribuibles a factores de riesgo es el siguiente: el 30-40 % se debe al colesterol sérico, el 20-25 % al tabaco, el 20-25 % a la hipertensión arterial y el 10-20 % a la inactividad física. Tercero, la medida en que una CC afecta a la frecuencia de la enfermedad depende de la frecuencia relativa de las restantes CC pertenecientes a la misma CS en la población. Así, si en la CS I de una enfermedad, la CC “E” está presente en el 50 % de la población, y el resto de CC (A, B, C, D) están presentes de forma combinada en la población en un 20 %, entonces “E” contribuirá a que enfermen un $0,50 \times 0,20 = 10\%$ de los individuos de la población. Si, por el contrario, “E” sigue presente en el 50 % de la población, y el resto de CC de forma combinada en el 2 % de ella, entonces la presencia de “E” contribuirá a que enfermen un 1 % de los individuos. Del mismo modo, la fuerza o riesgo relativo (RR) de una CC depende de la frecuencia relativa de las restantes CC pertenecientes a la misma CS en la población. Es por ello que existe interacción entre las CC de una CS, que depende de la prevalencia relativa de dichas CC en la población. Se habla de sinergismo si el efecto conjunto es superior a la suma de los efectos por separado, y de antagonismo si es al revés. Es decir, hay interacción (bajo un modelo aditivo) si el riesgo atribuible (RA) a la exposición conjunta de dos CC (A y B) es superior a la suma de sus RA por separado: $RA(A + B) > RA(A) + RA(B)$. Y antagonismo, si es al revés. Por último, los períodos de inducción no son constantes para las enfermedades, sino específicos de cada CC. Las CC actuando al inicio de la secuencia causal tienen un período de inducción largo; las CC finales, breve. Además, no hay tiempos de inducción constantes para una enfermedad en relación con sus CC. Dosis pequeñas requieren un conjunto de CC para completar una CS mayor que las dosis grandes. Estas tienen un período de inducción más corto al precisar menos CC para completar una CS.

3.2. Modelo probabilístico

En el modelo probabilístico o estadístico, causa se define como la condición o característica antecedente que aumenta la probabilidad de que se produzca un efecto. (12) El modelo probabilístico no excluye la causación necesaria y suficiente, que son simplemente extremos de esa definición: una causa necesaria es aquella que tiene que estar presente para que la probabilidad de la enfermedad sea mayor de 0 y una causa suficiente es aquella que aumenta esa probabilidad hasta 1. El modelo probabilístico es, por tanto, más inclusivo que el determinista. A diferencia del modelo determinista, el probabilístico no pretende dar cuenta del conjunto de acciones, eventos o estados que conducen a un desenlace concreto, por lo que no

requiere un conocimiento detallado de los mecanismos de acción del efecto de la exposición en el desenlace. (14)

El modelo probabilístico, como el modelo determinista modificado, contempla la posibilidad de que existan causas desconocidas que afectan a la ocurrencia de la enfermedad, interaccionando unas con otras, pero no depende de asunciones relativas a estas posibles relaciones. La probabilidad es un continuo entre 0 y 1, y el modelo probabilístico admite ese rango de posibles probabilidades para cada individuo, ya que no hay razón para pensar que una causa aumenta la probabilidad de desarrollar un efecto en la misma medida para cada individuo. De modo que cuando hablamos de causalidad, nos referimos a los efectos causales promedios en la población. Más aún, cabe la posibilidad de que incluso si conociésemos todas las causas de una enfermedad y dispusiésemos de información sobre las mismas para un individuo concreto, fuese imposible determinar con certeza si ese individuo desarrollará o no la enfermedad, porque se trate de un fenómeno inherentemente estocástico, no determinista, en parte sujeto al azar y regido por las leyes de la mecánica cuántica, de modo que solo se pueda determinar que ese sujeto tendrá una probabilidad X de desarrollar la enfermedad. Sea como sea el mundo a nivel de los individuos, estocástico (mecánica cuántica) o determinista (mecánica clásica), en epidemiología el interés se centra en el efecto causal promedio en la población diana de la que se ha obtenido la muestra para estudio, y tanto los estimadores de estos efectos poblacionales como sus intervalos de confianza son idénticos en ambos mundos.

3.3. Método contrafáctico

El método contrafáctico, orientado a la identificación de causas, incorpora una condición (*ceteris paribus*) a cualquiera de los modelos previos que permite distinguir la relación causal de la mera correlación, sea esta relación determinista o probabilística. En el modelo determinista, la hipótesis de investigación plantea que la causa (suficiente) producirá el efecto (una exposición producirá la enfermedad o un tratamiento la curación); mientras que en el modelo probabilístico la hipótesis dice que la causa aumentará la probabilidad del efecto. En la práctica, el modelo determinista modificado, ante la imposibilidad de identificar causas suficientes (por desconocimiento de algunas causas componentes), asume el planteamiento probabilístico para incorporar el azar debido a esa ignorancia. Así, el método contrafáctico establece que un factor antecedente es una causa de un evento si la probabilidad del evento es mayor en presencia del factor que en su ausencia, a condición de que el resto de condiciones se mantengan constantes. Este método es útil para orientar la investigación epidemiológica y la práctica de salud pública, ya

que la epidemiología persigue descubrir factores manipulables, que ofrezcan posibilidades para la intervención preventiva. Aunque no es posible cumplir la condición *ceteris paribus* en el mundo real, ya que no podemos observar a los mismos sujetos en la misma época expuestos y no expuestos a un factor determinado; en la práctica, los ensayos aleatorizados y controlados representan una aproximación adecuada a la misma que asegura la intercambiabilidad de los grupos comparados, permitiendo inferir la causalidad de las relaciones observadas (ver capítulo de estudios de intervención). De esta forma, los ensayos aleatorizados y controlados permiten identificar efectos causales incluso si los mecanismos que explican tales efectos son desconocidos.

Cuando no es posible realizar estudios de intervención, por razones de índole práctica (factibilidad, urgencia) o ética, podemos recurrir a los estudios observacionales. (15) En estos la intercambiabilidad se da cuando la probabilidad condicional de estar expuesto al factor de interés depende únicamente de las covariables medidas. (16) Como nunca es posible asegurar con certeza la intercambiabilidad en este tipo de estudios, para inferir causalidad es necesario, por un lado, conocimiento experto sustantivo (siempre falible) de la estructura causal del fenómeno de interés que oriente el análisis estadístico ulterior de las relaciones observadas en los datos crudos (ver capítulo de gráficos causales), y, por otro lado, recurrir a una serie de criterios de causalidad, que revisaremos más adelante.

El método contrafáctico requiere, también cuando recurrimos a estudios observacionales, plantear las preguntas causales en términos de un contraste de hipótesis sobre intervenciones realizadas durante un período de tiempo preestablecido, que es el objetivo de la inferencia, tal como se especificaría en el protocolo de un hipotético estudio de intervención. (17) Para ello, hay que definir el diseño del hipotético ensayo que serviría para estimar el efecto causal de interés, identificar una base de datos observacional con información suficiente para emular el ensayo y, de esa forma, estimar el efecto causal de interés usando datos observacionales. Para este modelo, las preguntas causales que no pueden ser trasladadas a un estudio de intervención, real o hipotético, no están bien definidas, bien porque haya más de una versión de la intervención en cuestión, bien por presencia de interferencia entre los sujetos del estudio. (18)

De lo anterior se desprende que la principal limitación del método contrafáctico es que solo es aplicable a factores proximales, mediadores manipulables con pequeños efectos potenciales, aislados de la complejidad en que se producen los cambios en la realidad, excluyendo del foco a los factores sociales, cruciales en salud pública, que se caracterizan por ser distales y complejos, y porque la interferencia (por interacción entre humanos) forma parte (integrada de forma indisoluble) de su constructo. (19) Además, los factores beneficiosos para la salud, como

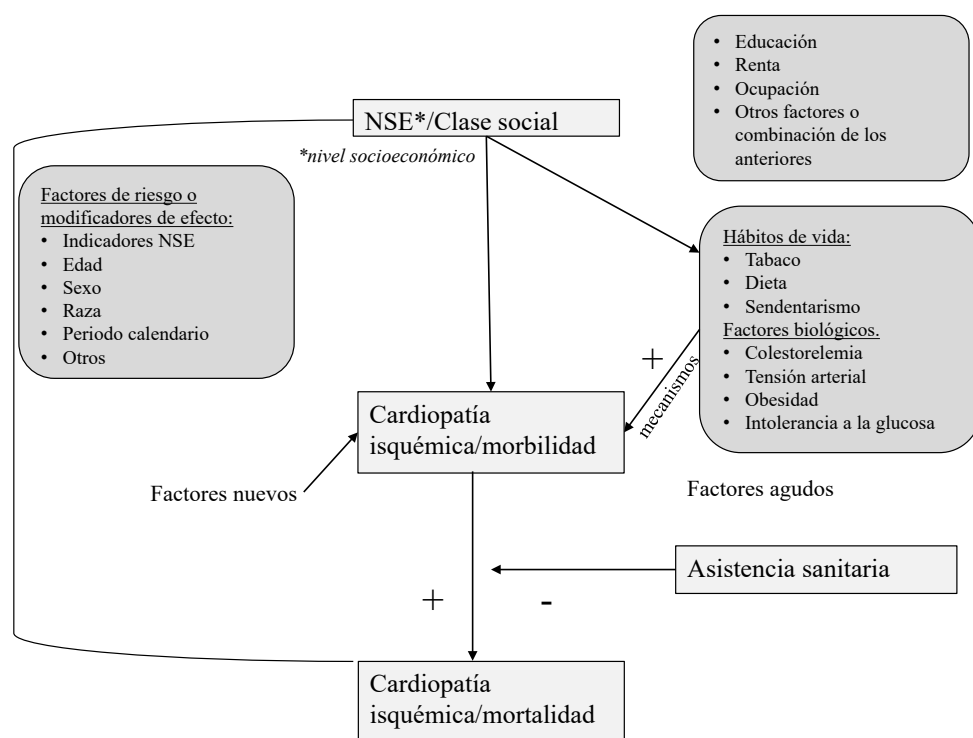
medicamentos, vacunas o actividad física, se acomodan mejor a este modelo que los factores perjudiciales, como la obesidad o el tabaquismo. Esto puede llevarnos a poner el foco en intervenciones concretas, como la terapia sustitutiva para dejar de fumar o una dieta hipocalórica para reducir la grasa corporal, meros parches para problemas de salud pública como la obesidad y el tabaquismo, en lugar de centrarnos en las causas distales, los complejos factores sociales que están detrás de esos problemas, desestimando las intervenciones que van a la raíz de los mismos, con mayor impacto potencial en la salud pública y capacidad transformadora del sistema que los alimenta.

3.4. Causación a múltiples niveles: Modelo de red

A pesar de las bondades de los modelos anteriores, se dispone actualmente de escasa información sobre las varias causas de muchas enfermedades. De ahí el interés de otros modelos causales más clásicos. En epidemiología, muchos de los factores causales que se estudian no son mecanismos bioquímicos o moleculares de las enfermedades, aunque estos brinden verosimilitud a las relaciones epidemiológicas. Sin embargo, tanto unos como otros son causas. El modelo multicausal de red acomoda de forma secuencial las causas distales y proximales de la enfermedad en un solo esquema. Por lo general, una enfermedad tiene varias causas que, conjunta o independientemente, dan lugar a su aparición. De igual manera, una sola causa puede contribuir a la aparición de varias enfermedades diferentes. Las causas de las enfermedades pueden dar origen a estas de forma indirecta —es decir, a través de la producción de otras causas (mecanismos intermediadores) de enfermedad— o directa. En realidad, casi siempre hay indicios de causas intermediadoras; por eso la relación de causalidad está mejor caracterizada por el grado de proximidad espaciotemporal con el desenlace que por su carácter directo o indirecto. De hecho, el que una asociación causal se considere directa o no depende del estado actual del conocimiento de sus mecanismos de acción, y no tiene implicaciones para la prevención. La llamada “red causal” es un modelo que ilustra de forma secuencial las relaciones entre causas y enfermedad. La figura 3 muestra simplificada la red causal o modelo teórico de producción de la cardiopatía isquémica. Las causas que ocupan en la red los niveles jerárquicos más altos, a menudo características sociales, determinantes a su vez de estilos de vida o factores relacionados, son del máximo interés desde el punto de vista epidemiológico y preventivo. El modelo teórico es un modelo reduccionista (pueden existir varias redes causales para una enfermedad dada) que ayuda a identificar los elementos implicados (principales, confusores, modificadores de efecto, etc.) en estudios concretos y a comprender que la prevención de la enfermedad depende de la eliminación de un elemento troncal de la red. Este modelo está relacionado con el modelo determinista modi-

ficado. Cada uno de los eslabones en la cadena causal pueden ser componentes de una CS. Sin embargo, cada una de las CC tiene sus propias redes causales, y una enfermedad determinada puede resultar de varias redes de CS, que pueden tener o no componentes en común. Por último, recientemente se han desarrollado en epidemiología, procedentes de las ciencias sociales, modelos gráficos de causalidad que, mediante la representación de las relaciones entre variables, permiten plasmar conceptos epidemiológicos como causalidad, causalidad inversa y asociación no causal, e identificar la presencia de confusión y sesgos de selección (ver capítulo de gráficos causales).

Figura 3. Cardiopatía isquémica. Modelo teórico de producción.



Fuente: elaboración propia.

4. Criterios de causalidad

En los estudios observacionales, para inferir causalidad es necesario recurrir a argumentos ajenos a los datos del propio estudio, puesto que no es posible asegurar la intercambiabilidad de los grupos comparados (equilibrio en la distribución de factores pronósticos). Los ensayos aleatorizados y controlados con ocultamiento de la secuencia de asignación sí que permiten inferir causalidad cuando el número

de sujetos aleatorizados es suficiente para garantizar la intercambiabilidad, siempre que exista correspondencia en las intervenciones aplicadas y un cumplimiento estricto del protocolo, con nulas o escasas pérdidas durante el seguimiento y ausencia de contaminación entre grupos (ver capítulo de estudios de intervención). En la práctica, resulta difícil que se cumplan todas estas condiciones, y en un único ensayo siempre cabe la posibilidad de que los resultados obtenidos sean debidos al azar (error de tipo I), por lo que los criterios de causalidad, en particular los de concordancia y verosimilitud, también contribuyen a apoyar la inferencia causal en los estudios de intervención.

Cuando las principales enfermedades eran infecciosas, la cuestión se centraba en saber qué criterios era necesarios para probar que un organismo causaba la enfermedad. En 1840, Henle propuso unos postulados de causación, expandidos por Koch en la década de 1880. (20) Los postulados de Henle-Koch requerían que el organismo fuera recuperado de cada caso de enfermedad y cultivado *in vitro*, que su inoculación en otro huésped causara la enfermedad y que fuera aislado de nuevo en este último. La validez del paradigma de Koch ha sido demostrada para varias enfermedades infecciosas. Sin embargo, un factor por sí solo rara vez es suficiente para causar enfermedad, incluso cuando se trata de una causa necesaria, y esto vale también para enfermedades de origen infeccioso como la tuberculosis, cuyo desarrollo depende no solo de la presencia de *Mycobacterium tuberculosis*, sino también de la dosis infectiva y de otros factores como el estado nutricional e inmunitario del sujeto expuesto al microorganismo o las condiciones ambientales. (21) Conforme las enfermedades no infecciosas ganaron importancia a mediados del siglo XX, MacMahon, Pugh e Ipsen introdujeron el concepto de “red de causalidad” para enfatizar la importancia de las múltiples causas de la enfermedad. (22) En 1954, Edward Cuyler Hammond enunció una serie de criterios, posteriormente discutidos y ampliados por Yerushalmy y Palmer en 1959, Sartwell y otros. En 1964, un comité de expertos nombrado por el Surgeon General de Estados Unidos desarrolló directrices para evaluar la evidencia sobre la relación entre el tabaco y el cáncer de pulmón. (23) Entre sus revisiones posteriores, por ser las más conocidas y usadas en epidemiología, comentamos de forma detallada las guías o criterios que sir Arthur Bradford Hill publicó en 1965 para valorar la posible causalidad de una asociación. (24)

4.1. Criterios de Hill

Los criterios o estándares se dirigen a la evaluación de la evidencia disponible sobre la relación entre dos factores: generalmente, la exposición a un agente y la ocurrencia de una enfermedad. No comentaremos los criterios en el orden en que Hill los enumeró (tabla 1), porque no estableció una jerarquía entre ellos. Según

dijo, “ninguno de los nueve puntos de vista es incuestionable para aceptar o rechazar la causalidad”. Además, los comentamos más allá de lo descrito por Hill, haciendo referencias a su moderna valoración. Muchos de los criterios se corresponden con los enunciados antes y después por otros autores. Aunque muchos autores (y el propio Hill) recomiendan usar los criterios causales solo ante una relación estadísticamente significativa, matizan esta cuestión. Por ejemplo, si el tamaño muestral es pequeño, hay que interpretar una diferencia sustantiva no significativa como “no probada” y no como “no real o no causal”. En definitiva, dice Hill, cuando la diferencia observada es muy grande o muy constante, sería absurdo atribuirla al azar y carece de importancia un test de significación; y si es muy pequeña, no interesa si estadísticamente es significativa o no. Pero entre estos dos extremos, los test estadísticos ayudan en la interpretación.

Tabla 1. Criterios de causalidad de A. B. Hill, 1965

Criterio	Descripción
1. Fuerza	Riesgo relativo grande.
2. Concordancia	Asociación observada repetidamente por varias personas, en sitios, circunstancias y épocas diferentes.
3. Especificidad	Una causa lleva a un solo efecto.
4. Temporalidad	La causa precede al efecto.
5. Gradiente biológico	La magnitud de la enfermedad aumenta con la magnitud de la exposición a la causa.
6. Verosimilitud	La asociación tiene sentido de acuerdo al conocimiento biológico del momento.
7. Coherencia	Ausencia de conflicto con la historia natural y biológica de la enfermedad.
8. Experimento	La reducción de la exposición a la causa se asocia a una disminución de la enfermedad.
9. Analogía	Relación causa-efecto ya establecida para un agente-enfermedad similares.

Precedencia temporal

La exposición al factor presuntamente causal debe preceder en el tiempo a la aparición del efecto, pues podría ocurrir que fuera la consecuencia del “efecto” (causalidad inversa); así, muchos estilos de vida (p. ej., la dieta) cambian tras (y a consecuencia de) las etapas iniciales de la enfermedad que presuntamente causan, sobre todo si es una enfermedad crónica de instauración lenta. No obstante, que la “causa”

siga al efecto no significa que la “causa” no pueda causar ese efecto en los casos en que le precede (la dieta desequilibrada puede causar enfermedad, aunque en una situación dada se altere tras los primeros síntomas de esa enfermedad).

Un refinamiento útil de este criterio es considerar que el tiempo en que la causa putativa precede al efecto es relevante, es decir, compatible con el período de inducción de la enfermedad o consistente con un mecanismo biológico. No todos los estudios epidemiológicos tienen la misma capacidad para poner de manifiesto este criterio, que solo es obvio en los estudios de intervención o con variables fijas. Si el estudio es de cohortes prospectivo, el criterio es más fácilmente comprobable que si es de casos y controles retrospectivo. Si el estudio es transversal, el criterio queda garantizado solo si el factor de exposición es una característica fija (p. ej., el grupo sanguíneo o la fecha de nacimiento). No solo el tipo de diseño, sino también la medida de frecuencia elegida tiene diferente eficacia para demostrar la secuencia cronológica: es mejor la incidencia que la prevalencia, y la prevalencia de punto a la de sobrevida.

Este es el único criterio considerado por algunos autores como *sine qua non*. Sin embargo, el propio Hill no lo consideró imprescindible, probablemente por la dificultad de su constatación en muchas ocasiones. Ser taxativo en la exigencia de este criterio llevaría a despreciar muchos estudios observacionales. Además, el período de inducción de la última CC de una CS cuyos componentes actúen secuencialmente (algunos cánceres o enfermedades cardiovasculares) es prácticamente nulo, es decir, la causa sería prácticamente simultánea al efecto. Por ejemplo, el aumento de mortalidad por enfermedades respiratorias en ancianos que sigue, con una diferencia de horas o días, al aumento de concentración de partículas por contaminación ambiental, es perfectamente compatible con periodo de inducción prácticamente nulo, ya que es una de las últimas CC en actuar. (25)

Fuerza

También denominado magnitud o tamaño del efecto, el primer criterio que Hill puso en su lista, es la intensidad de la asociación entre el factor y la enfermedad. En epidemiología, la fuerza de la asociación es usualmente medida por el grado en que el RR o la *odds ratio* se separan de la unidad, sea por encima de 1 (en el caso de exposiciones que causan enfermedad) o por debajo de 1 (en el caso de exposiciones preventivas). Pero también puede estimarse por una diferencia de riesgos, de medias o de proporciones, o por un coeficiente de regresión. La medida relativa del riesgo es más informativa que la diferencia absoluta para valorar la fuerza de la asociación: los números relativos son más elocuentes en relación con la etiología. Así, en la relación del tabaco con el cáncer de pulmón, un RR de 8 (ratio de tasas 0,57-0,07/1000 personas y año en fumadores de 1-14 cigarrillos/día

frente a no fumadores) es más elocuente que una diferencia de tasas de 0,50/1000 personas y año. Asimismo, una diferencia de 2,2/1000 de mortalidad entre dos grupos (fumadores y no fumadores) es menos probable que se deba a un error de medición si se presenta entre las tasas 2,27 y 0,07 (cáncer de pulmón, $RR = 32$) que entre las tasas 9,9 y 7,3 (enfermedad cardiovascular, $RR = 1,4$).

La regla de interpretación de este criterio es que cuanto mayor es la fuerza, más verosímil es la hipótesis causal (o más “probablemente” causal es, se suele decir). Un RR muy alto (p. ej., > 10 , como para el tabaco y el cáncer de pulmón) solo puede deberse a una variable de confusión si es mucho más frecuente en los expuestos que en los no expuestos y se asocia fuertemente con la enfermedad. Los confusores rara vez dan cuenta de una diferencia tan considerable. Además, para Hill, un RR muy alto requiere un factor tan íntimamente ligado al factor estudiado que sería fácilmente detectable. Inversamente, RR próximos o inferiores a 2 (como en la relación tabaco-trombosis coronaria) inducen a pensar en factores de confusión que podrían explicar tales efectos.

Se requiere experiencia en el área de investigación para apreciar cuándo un RR u otra medida del tamaño del efecto es realmente grande, importante para constituir una buena evidencia de causalidad. El propio Hill reconoce que una asociación débil puede ser causal. Así lo es la relativamente débil relación entre consumo de cigarrillos y enfermedad cardiovascular o la del tabaquismo pasivo y el cáncer de pulmón. En el contexto de las principales enfermedades crónicas, autores como Brownson o Hill hablan de RR débiles si son inferiores a 2; moderados, entre 2 y 4, y fuertes si son superiores a 4. (26) En cambio, Potischman y Weed han propuesto recientemente, en el campo de la epidemiología nutricional, que RR estadísticamente significativos por encima de 1,2 o por debajo de 0,8 son valorables en el juicio causal, dada la rareza de encontrar RR mayores de 1,2 o menores de 0,8, por lo que un RR de 1,4-1,5 (o 0,5-0,6) sería fuerte. (27) Otros autores, como Kassirer y Angell, por el contrario, son más exigentes y solicitan RR de al menos 4 o 5 para empezar a considerar que una relación sea causal. (28) Sin embargo, exigir valores concretos al RR equivale a olvidar que los criterios no son reglas lógicas, autosuficientes. La valoración del tamaño del efecto es una cuestión de juicio informado. A este juicio puede ayudar el examinar cuán consistentemente grande es el efecto a lo largo de una serie de estudios. En el campo de la salud, donde RR por encima de 4 son muy raros, la experiencia nos muestra que cuando encontramos un RR alto en un estudio, lo habitual es que el efecto disminuya notablemente o se desvanezca al llevar a cabo nuevos estudios. (29) Por ej., los primeros estudios que analizaron el efecto de la proteína C reactiva en la enfermedad coronaria detectaron un RR en torno a 4, un efecto que, combinado con el de estudios posteriores, se ha reducido hasta un RR de 1,6. (30) Asimismo, el supuesto efecto preventivo

del sulfato de magnesio en pacientes con sospecha de infarto de miocardio detectado en los primeros ensayos llevados a cabo para examinar la cuestión ($RR = 0,5$), se desvaneció al realizar nuevos ensayos, con muestras más grandes, que mostraron un posible aumento de la mortalidad en los pacientes tratados. (31) Tales errores, especialmente con resultados obtenidos en estudios de pequeño tamaño muestral, pueden ser debidos al azar combinado con posibles sesgos en el diseño, ejecución o publicación de los estudios. También una relación fuerte puede ser no causal debido a factores de confusión no controlados. Es el caso del síndrome de Down y el orden de nacimiento confundido por la edad materna. Por supuesto, una vez que el confusor está identificado, su control reduce la asociación.

Otra limitación del RR (u otra medida de la intensidad de la asociación) es que su magnitud depende de la prevalencia de las otras causas componentes necesarias para completar la causa suficiente o mecanismo causal del que forma parte el factor en cuestión. Cuanto más frecuentes son las causas complementarias de un factor, mayor es la fuerza de este, y viceversa. De este modo, el RR es diferente entre poblaciones con distinta distribución de factores de riesgo: el RR de enfermedad coronaria por hipercolesterolemia en dos poblaciones con similar frecuencia de esta será mayor en aquella que tenga mayor proporción de fumadores o hipertensos (las tres variables forman parte de la misma causa suficiente de la enfermedad, digamos la placa de ateroma).

En conclusión, una asociación fuerte solo sirve ni más ni menos que para descartar la hipótesis de que la asociación se debe enteramente a un confusor u otra fuente de sesgo modesto. Pero cuanto más específico es el constituyente de una causa, más cerca del evento biológico está, o más específica o precisa es la medida de la exposición, más fuerte es la asociación. Cuando —como es frecuente— la variable de exposición medida es una estimación muy indirecta del verdadero factor biológico de interés, la asociación es débil (p. ej., medir un alimento en lugar de una vitamina). (15) Cuando el factor evaluado es un estimador más próximo del agente biológico verdadero, el RR será mucho más alto. No obstante, el RR también será más alto con factores distales que ejercen su acción a través de varios mecanismos causales intermediarios, por ej., el patrón de dieta mediterráneo o la clase social. Por estas razones, un RR fuerte es más verosímilmente causal; pero la experiencia nos enseña que cuando se publican los primeros resultados sobre nuevos factores de riesgo o preventivos con RR altos, lo prudente es esperar a que nuevos estudios, con muestras grandes y bien diseñados (bajo riesgo de sesgos y buen control de la confusión), corroboren o descarten los efectos observados inicialmente. Cuando los resultados se replican, un RR alto habla en favor de la hipótesis de causalidad.

Concordancia (Consistency)

No basta un resultado aislado para refutar o corroborar en la práctica una hipótesis causal. Es decir, aunque la evidencia circunstancial puede apoyar positiva o negativamente una hipótesis, ni el mejor y mayor estudio aislado descarta por completo toda fuente de error. Los estudios aislados son raramente definitivos, salvo, si acaso, cuando se trata de estudios de intervención bien diseñados y ejecutados, y con suficiente tamaño muestral para que la probabilidad de error aleatorio resulte despreciable, como el *Women's Health Initiative*, (32) que observó un mayor riesgo cardiovascular en mujeres postmenopáusicas que recibieron tratamiento hormonal sustitutorio, o los ensayos que demostraron la ausencia de efecto quimiopreventivo del β -caroteno para el cáncer de pulmón, (33) contradiciendo en ambos casos los resultados de los estudios observacionales previos.

En epidemiología no es posible alcanzar el grado de control del laboratorio y, por lo tanto, el sustituto de la replicación exacta es la repetición en condiciones similares; basta que se reproduzcan razonablemente las condiciones relevantes del estudio (sujetos comparables, misma variable de resultado, etc.) para comprobar si el efecto es similar, es decir, si hay concordancia. Hallazgos similares obtenidos en diferentes poblaciones, bajo circunstancias y con métodos distintos, y por diferentes investigadores, tienen un peso mayor para apoyar la causalidad. Es poco probable que estudios diferentes concluyan una misma cosa si esta no es cierta; por ejemplo, el juicio causal sobre la relación entre el consumo de cigarrillos y el riesgo de enfermedad coronaria se fortaleció cuando se realizaron un gran número de estudios de casos y controles y de cohortes en los últimos 40 años en un amplio rango de poblaciones de culturas distintas, observando a millones de personas, y todos ellos mostraron un riesgo aumentado. La observación de un aumento del riesgo cardiovascular también entre los fumadores pasivos, con una gran concordancia entre estudios, refuerza aún más la hipótesis causal. (34)

Hay dos tipos de replicación: la primera es la mera repetición, con ánimo de verificar un resultado, y la segunda es la que varía las circunstancias o métodos para ver si la hipótesis resiste nuevas pruebas independientes, duras, severas: por ejemplo, con medición más precisa (técnicas más fiables) de la exposición o de la enfermedad, poblaciones de distintas características sociodemográficas, períodos de tiempo variables, o consideración de otros factores causales o de confusión, etc. Esta segunda es la más rica y útil para la inferencia causal. Por ello, la comprobación de la hipótesis de Barker sobre los orígenes fetales de la enfermedad coronaria ha sido acusada de verificacionismo, lo cual no quiere decir que no sea cierta: para algunos solo se han buscado pruebas a su favor, pero no se ha contrastado duramente, para ver si surgen contraejemplos (controlando el nivel socioeconómico a lo largo de la vida, o estudiando gemelos, o buscando mecanismos verosímiles).

(35) En filosofía de la ciencia, la refutación es más fuerte que la verificación. Las pruebas afirmativas son más fáciles de buscar y encontrar (sobre todo en los primeros estudios), y mientras que un conjunto finito de observaciones pueden refutar una hipótesis, bajo ninguna condición podrían verificarla al concordar con esta (siempre cabrá la posibilidad de que existan casos inobservados que pudieran contradecirla). La cuestión de cuándo una relación es concordante es materia de juicio; aunque, como en otros criterios, no es un asunto arbitrario. Además, que no haya un criterio nítido no significa que no haya una idea reguladora, una regla que nos oriente y que facilite el acuerdo. Una regla razonable puede ser la siguiente: para ser concordantes, los hallazgos obtenidos en estudios comparables y correctamente realizados deben ir en la misma dirección (RR por encima o por debajo de 1). Aunque los resultados pueden diferir en el tamaño del efecto al variar las características de los estudios (rango de edad, prevalencia de otras CC de la misma CS, etc.), y esto puede tener trascendencia práctica, los hallazgos de magnitud similar son más concordantes.

Aunque la concordancia refuerza la verosimilitud de la hipótesis causal, su ausencia no descarta por completo una relación causal, porque algunos efectos ocurren solo bajo circunstancias inusuales o en algunas personas especialmente susceptibles a un factor de riesgo. Más precisamente, el efecto de un agente causal no puede producirse a menos que actúen antes, después o al mismo tiempo el resto de CC que se complementan hasta producir una CS. Estas condiciones no son siempre satisfechas. Así, el contacto con el virus VIH puede causar infección, pero no siempre lo hace, ya que el virus tiene que estar presente en una concentración determinada y en una zona permeable a su entrada (p. ej., zona de la piel con pérdida de integridad), o porque la persona en cuestión presenta una variante genética que la hace inmune al virus. Asimismo, en ocasiones ocurren acontecimientos singulares, como el exceso de cáncer de pulmón y senos nasales en los trabajadores de una refinería de níquel en Gales del Sur a principios de siglo, donde los niveles de níquel eran extraordinariamente altos. Aparte de esto, son frecuentes los resultados dispares al repetir estudios. Ante resultados discordantes, no se trata de comparar el número de estudios con resultado positivo con el de los negativos, sino que cada estudio debe ser evaluado según su calidad, sesgos y otros criterios que veremos más adelante, y además hay que ponderarlos, siquiera sea por juicio cualitativo, y descartar la presencia de sesgo de publicación. Los resultados negativos falsos pueden ocurrir por un tamaño muestral insuficiente, sesgos o factores de confusión en estudios débiles, mal diseñados o pequeños. Un estudio que falle a la hora de encontrar una asociación, pero que es limitado en metodología y tamaño, de modo que no puede detectar un efecto existente por falta de potencia estadística o presencia de sesgos, es de poco valor para contradecir un resultado, aunque este sea novedoso y controvertido. A veces, los resultados aparentemente

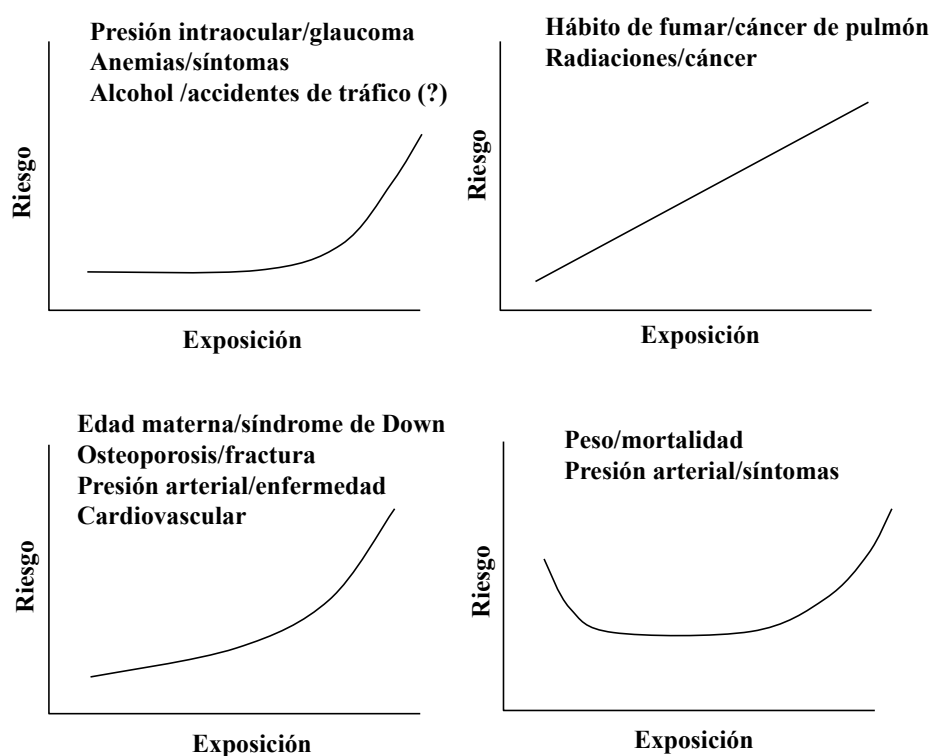
contradictorios pueden explicarse por la presencia de modificación de efecto, es decir, porque la exposición tiene efectos distintos en diversos sectores de la población. En cualquier caso, si los hallazgos de varios estudios son discordantes, la discrepancia tiene que ser explicada. Por último, si la asociación se repite, podría deberse a un sesgo que ocurra siempre en la misma dirección. En consecuencia, ni la ausencia ni la presencia de concordancia garantiza la causalidad. Por último, la concordancia está muy relacionada con la generalización (validez externa o aplicabilidad de los resultados a otras poblaciones). La repetición en diversas circunstancias es el mejor test de generalización, y la concordancia su mejor criterio. Por ejemplo, el estudio de Framingham para examinar las causas de la enfermedad coronaria se diseñó originalmente en individuos de edad media, de clase media, en una muestra sistemática en la pequeña ciudad norteamericana de Framingham; sin embargo, sus resultados (p. ej., la hipercolesterolemia como factor de riesgo cardiovascular) se aplican en otras partes del mundo, y no solo porque la hipótesis es universal por la constancia de las relaciones biológicas subyacentes (y, por ende, de los patrones de enfermedad) entre diferentes individuos, sino también porque se han obtenido resultados similares al volver a examinar esas relaciones o hipótesis en otras poblaciones, otras circunstancias y con otros métodos de estudio. (36) Además, la generalización no es solo una manera de examinar la aplicabilidad de los resultados, sino además un método de enriquecer o elaborar la hipótesis causal, de corroborarla más fuertemente, pues una hipótesis más general ha probado su temple en más circunstancias.

Gradiente biológico o relación dosis-respuesta

Para algunos autores es una variante del criterio de magnitud. Nos dice que, si un factor es causal, cuanto mayor es la exposición al factor, mayor es el riesgo de enfermar. Es decir, hay una asociación positiva (o negativa si el factor es preventivo) de carácter progresivo (o regresivo) entre los niveles de exposición al factor y la magnitud de su efecto. Si es posible, las variables predictoras deben medirse de forma continua o mediante varias categorías para estudiar la relación dosis-respuesta. La asociación entre fumar cigarrillos y cáncer de pulmón es un ejemplo de esta relación: los fumadores moderados tienen tasas de cáncer más elevadas que los no fumadores, y los grandes fumadores presentan tasas aún mayores. (37) También la relación de las radiaciones con el cáncer parece ser aproximadamente lineal a lo largo de todo el intervalo de exposición. Sin embargo, otras asociaciones producen relación curvilínea (p. ej., osteoporosis/fractura, presión arterial/enfermedad cardiovascular), y otras son en forma de “U” o “J” (p. ej., la relación peso/mortalidad) (figura 4). Hay relaciones causales que no cumplen este criterio. Algunos efectos solo se producen por encima de un umbral de exposición, como

en el caso de las relaciones presión intraocular/glaucoma, o dietilestilbestrol y cáncer de vagina. Tampoco las asociaciones que muestren relación dosis-respuesta son necesariamente causales, pues pueden reflejar únicamente el efecto de un confusor no controlado; por ejemplo, la relación dosis-respuesta entre el tabaco y la cirrosis hepática se explica por la fuerte asociación monotónica entre el tabaco y el consumo de alcohol.

Figura 4. Relación dosis-respuesta en cuatro situaciones.



Fuente: Rose G. *La estrategia de la medicina preventiva*. Masson, 1994.

Verosimilitud

El apoyo que se demanda a una hipótesis causal no consiste solamente en datos que corroboren sus implicaciones (convalidación empírica); puede venir también de hipótesis o teorías más amplias que tengan apoyo empírico independiente. Si se añade esta convalidación teórica, la explicación es más profunda, más satisfactoria. En una palabra, no nos conformamos con correlaciones concordantes, universales, explicativas de hechos o problemas; queremos saber el mecanismo subyacente a las mismas.

Un mecanismo biológicamente verosímil, conocido o postulado, debe explicar por qué la exposición altera el riesgo de desarrollar la enfermedad. No obstante, dado que su consideración depende del estado actual del conocimiento, su falta no

descarta la causalidad. Un ejemplo es la asociación entre el tabaquismo y el cáncer de cérvix. Al principio le faltaba verosimilitud biológica y se creyó poco en ella, pero luego se identificaron componentes del humo del tabaco en el moco cervical y en otros líquidos corporales de las fumadoras. De hecho, la observación aparentemente no verosímil alentó la búsqueda de mecanismos. (38)

Por otra parte, su presencia no garantiza causalidad. Un mecanismo aparentemente verosímil para la asociación puede luego no demostrarse, y la asociación puede estar sesgada. Además, una hipótesis epidemiológica puede ser compatible con varias teorías mecanicistas. Sin embargo, una teoría mecanicista es preferible a una que no considera el mecanismo, por ser lógicamente más fuerte (implica a esta), más rica (mayor contenido), metodológicamente más exigente (da más numerosas y precisas oportunidades de contraste), y suministrar una explicación más profunda (detalla un nivel de la realidad distinto, más básico). Un ejemplo es la teoría oxidativa de las lipoproteínas de baja densidad como mecanismo para explicar la relación entre colesterol sérico y cardiopatía isquémica. (39)

La ausencia de este criterio se da en las llamadas relaciones o hipótesis de “caja negra”, que ignoran los mecanismos de la relación observada. Las relaciones de caja negra pueden ser valiosas (hay muchos ejemplos de ello); además, son más generales, sencillas, precisas y seguras. Pero buscar sus mecanismos (“iluminar el interior de la caja”) puede lograr un mejor conocimiento del problema y, por lo tanto, mejorar sus posibilidades de control. Aunque Snow no conocía la implicación de un microorganismo en la epidemia de cólera de Londres en el siglo XIX, ni lo necesitó para controlar la epidemia, (40) el conocimiento microbiológico (posterior) mejoró las medidas de prevención y tratamiento de este mal. Además, que se busquen los microhechos biológicos que subyacen a los macrohechos epidemiológicos es independiente y compatible con que se busquen sus mecanismos sociales, las causas de las causas. No hay antagonismo entre ecoepidemiología, epidemiología social, epidemiología de estilos de vida, epidemiología molecular y epidemiología genética. (10,41)

Por último, el criterio de verosimilitud está relacionado con el de concordancia o, mejor dicho, con el de repetibilidad. (8,41) Al fin y al cabo, considerar los mecanismos equivale a examinar si la hipótesis causal se mantiene y es coherente con circunstancias evidenciables biológicas, que son test más intensos y persuasivos de la hipótesis. Además, este criterio está relacionado con la generalización: considerar los mecanismos conocidos o postulados de una asociación ayuda a decidir si es aplicable a otros grupos de sujetos.

Por todo lo anterior, la verosimilitud es un criterio de gran relevancia. Sin embargo, la mera postulación de un mecanismo causal sin una contrastación posterior del mismo tiene escaso valor, ya que en la literatura científica podemos en-

contrar hipótesis de mecanismos causales a las que apelar, por analogía, con casi cualquier posible relación causal que analicemos, especialmente en campos como la epidemiología nutricional o genética (29). A efectos de dilucidar la verosimilitud de tales hipótesis, en los últimos años se han producido importantes avances metodológicos para el análisis de mediadores causales en el marco del método contrafáctico. (42,43)

Especificidad y analogía

La especificidad postula que una causa conduce a un efecto único, no a efectos múltiples. Sin embargo, muchos factores aislados suelen tener muchos efectos (p. ej., el tabaco o la actividad física). En todo caso, cierta especificidad en la magnitud de la asociación puede contribuir a reforzar una hipótesis causal (así, en la relación tabaco-cáncer de pulmón, el RR aumenta mucho más que en la del tabaco con otras enfermedades).

La analogía postula que asociaciones causales similares podrían producir enfermedades similares. Por ejemplo, si un medicamento puede causar malformaciones congénitas, quizá puede hacerlo otro. En campos como la epidemiología nutricional o la farmacoepidemiología es habitual encontrar alguna analogía en la literatura para casi cualquier desenlace de interés. (29) Este criterio, contemplado por Hill y MacMahon, forma parte del concepto global de verosimilitud, pues se refiere a que la relación es análoga a alguna otra relación bien establecida; y, más bien, es útil para formular hipótesis más elaboradas que para evaluarlas.

Ni la especificidad, por ser raramente observable en enfermedades no transmisibles, ni la analogía, precisamente por lo contrario, son criterios relevantes para establecer una relación causal.

Experimento

También se le denomina cesación de la exposición. Si un factor tiene un efecto, dicho efecto debería cesar cuando es removido el factor. Durante la epidemia de cólera de Londres a mediados del siglo XIX, John Snow estudió detenidamente un brote especialmente virulento de cólera en el barrio de *Soho* a finales de agosto y primeros de septiembre de 1854, observando que la mayoría de los casos fatales ocurrían en personas que vivían en las cercanías de la bomba de agua de la calle *Broad*. (40) Por indicación de Snow, cuya hipótesis era que el brote se había originado por contaminación del pozo con aguas fecales, las autoridades clausuraron el pozo retirando la manivela, lo que redujo drásticamente la incidencia y mortalidad por cólera. Un ejemplo reciente es el del proyecto *Women's Health Initiative*, que contrasta entre otros los efectos de una dieta baja en grasas sobre la incidencia del cáncer de mama. (44)

Este criterio no siempre puede obtenerse en poblaciones humanas, pero si el experimento se puede llevar a cabo, constituye la prueba más sólida de causalidad. El criterio de experimentación es un criterio o regla metodológica de severidad o robustez de la hipótesis. Pero incluso puede ser visto, más que como un criterio, como un test directo de la hipótesis causal.

Coherencia

Para Susser y otros autores, la coherencia incluye los criterios de analogía y versosimilitud. (41) La coherencia dice que la asociación no debe entrar en conflicto con lo que se sabe de la enfermedad; es decir, debe ser consistente con el conocimiento existente. Así, la asociación del cáncer de pulmón con el consumo de tabaco es coherente con el aumento temporal ocurrido en las dos variables desde principios del siglo XX, y con la diferencia sexual en la mortalidad. Contribuyen a la coherencia las alteraciones histopatológicas observadas en el epitelio bronquial de los fumadores y la presencia en el humo del tabaco de carcinógenos para la piel de los animales de laboratorio. Pero la no coherencia (distinguir la del conflicto) no debería ser una prueba en contra de la causalidad. Así, la ausencia de confirmación de la hipótesis en animales de laboratorio no va en contra de la relación en seres humanos. Y si el factor causa solo una pequeña proporción de la enfermedad total, la influencia abrumadora de otros factores puede hacer al patrón global incoherente.

Recapitulación

En resumen, el experimento, en condiciones ideales, puede considerarse un test más que un criterio de causalidad, y la precedencia temporal es el único criterio que puede considerarse *sine qua non*, ya que forma parte de la propia hipótesis causal, aunque no siempre es posible su constatación fehaciente en estudios observacionales. Del resto de criterios, los más relevantes son la verosimilitud y la concordancia. Por último, la fuerza de la asociación y su gradiente biológico son importantes en teoría, pero es raro encontrar efectos de gran magnitud, y las relaciones entre una exposición y su efecto no siempre son progresivas (o regresivas). Esto nos recuerda que si un criterio está presente refuerza el juicio de causalidad, pero su ausencia no lo niega.

4.2. Otros criterios de causalidad

En un estudio de intervención asociación es causalidad, sin consideración del azar, pero en la práctica es bastante raro que se den las condiciones ideales de un estudio de intervención en la investigación con seres humanos, por lo que los

criterios de Hill son útiles para inferir la causalidad de las asociaciones observadas tanto en estudios observacionales como de intervención. Sin embargo, los criterios de Hill no capturan todos los elementos útiles para la inferencia causal ni son de mucha ayuda ante la supuesta causalidad de una asociación observada en estudios con limitaciones metodológicas de importancia. (8,16,41)

Criterio de independencia (descarte de explicaciones alternativas)

Puede haber explicaciones de la realidad alternativas a la causal, como son la causalidad inversa, los sesgos de selección e información, los factores de confusión y el azar. Para muchos autores, estos criterios “negativos” (el descarte razonable de sesgos y factores de confusión apoya la causalidad) deben ser la primera consideración en la evaluación de una hipótesis causal. El orden de la valoración debe ser primero los sesgos y luego los factores de confusión, pues si hay sesgo grave, ninguna manipulación de los datos salvará el problema; la valoración del azar vendrá en tercer lugar.

Una concreción de la consideración de los factores de confusión es el llamado por algunos autores criterio de “independencia”. Este criterio es considerado en muchos estudios bajo la forma general de que la asociación parece no deberse enteramente al efecto de un factor de confusión. El *U.S. Pooling Project* de estudios observacionales prospectivos sobre determinantes de la enfermedad coronaria lo incluyó entre sus criterios de causalidad: un factor de riesgo es independiente si mantiene al menos un cierto grado de efecto una vez controlada (generalmente por métodos multivariados) la influencia de otros determinantes relevantes de la enfermedad. (45) Este criterio lo cumplían en 1978 el aumento de colesterol sérico, el consumo de tabaco y la hipertensión arterial (tras ajustar por edad y los otros factores predictores). Una vez descartadas de forma razonable las explicaciones alternativas (en los estudios observacionales no es posible garantizar por completo la ausencia de factores de confusión desconocidos), los criterios de Hill servirán para reforzar o debilitar la hipótesis causal. De todas formas, como plantean algunos autores (e.g., Susser), el criterio de independencia, referido a factores de confusión, puede subsumirse como refinamiento del criterio de fuerza (consideración del RR ajustado y no solo del crudo), o incluso en el de concordancia (posible variación del efecto, o no variación, en circunstancias o estudios diferentes que consideran los mismos u otros factores de confusión ajustados).

Poder explicativo

Este criterio tiene que ver con la audacia, originalidad, novedad, contenido informativo, rendimiento o poder predictivo o explicativo, interés e importancia de una hipótesis (causal). Son características interrelacionadas que sirven para evaluar

a priori una hipótesis causal, ayudando así al progreso de la ciencia si esa hipótesis resulta corroborada *a posteriori*.

Una hipótesis es “audaz” si es nueva, va más allá de la vigente, predice hechos previamente insospechados, o da respuesta a una pregunta no contestada todavía o que lo ha sido de forma contradictoria o sin replicación. MacMahon resaltaba que, en epidemiología como en otras ciencias, el progreso en la búsqueda de asociaciones causales se favorece con hipótesis nuevas o más específicas. (2) Popper distingue dos componentes de la audacia. (9) Audacia de predecir aspectos observables de la hipótesis hasta ahora pasados por alto, que son tan osados que corren el riesgo de ser falsos y, por consiguiente, es preciso intentar refutarlos. Por eso, si resisten los test (criterios), se admite provisionalmente la hipótesis: es muy improbable que el éxito de una teoría tan audaz, tan improbable, tan poderosa se deba a errores o sesgos. La segunda audacia es suponer la realidad escondida tras lo observable: postular los mecanismos o las leyes biológicas subyacentes a la hipótesis o relación epidemiológica, y se relaciona con la verosimilitud biológica.

La audacia, originalidad, novedad o interés de la hipótesis *a priori* se relacionan con su poder para predecir o explicar más hechos observables novedosos e interesantes que la hipótesis previa vigente, o que una rival. Un rico contenido informativo de la hipótesis puede contrastarse de forma independiente y, por lo tanto, más severa, más fértil y más corroborable. Pero para que se produzca progreso en el conocimiento, algunas de esas nuevas predicciones deben ser corroboradas de hecho, superar comprobaciones ingeniosas y rigurosas. Esas pruebas rigurosas presuponen un alto grado de contraste o contenido *a priori*. De ahí la importancia del criterio llamado por Susser “rendimiento predictivo”: nuevas predicciones que resulten exitosas, concepto tomado de Popper. (41)

Un ejemplo de teoría audaz es la de Ross sobre la aterosclerosis (1976), según la cual la placa de ateroma es la consecuencia de la respuesta local a la lesión endotelial. La lesión endotelial daba lugar a la agregación plaquetaria, y el factor de crecimiento derivado de las plaquetas originaba una proliferación y emigración del músculo liso que producía la placa de ateroma. Incorporaba, pues, la plaqueta como nuevo elemento a la teoría clásica y subrayaba el fondo inflamatorio de la proliferación. Posteriormente (1986, 1993), nuevos hallazgos cambiaron la plaqueta por el monocito como célula fundamental. (46) La teoría actual es una síntesis de las tres hipótesis iniciales: la inflamatoria de Virchow, la trombótica de Rokitsansky y la lipídica de Anitschkow. El hecho clave de la aterosclerosis, particularmente de su comienzo, es la acumulación de lipoproteínas; la trombosis es fundamental en la placa rota inestable y el infarto de miocardio, y la inflamación es evidente en el papel de monocitos, macrófagos y proliferación muscular lisa en la génesis del ateroma. La moderna teoría subsume, pues, a las anteriores, predice

los hechos previstos por ellas y, además, añade nuevas predicciones; es más rica en contenido (y verosímil tras contrastarse). El interés de una teoría depende tanto de la sorpresa como de la importancia. El interés teórico es el potencial para cambiar lo que los científicos creen sobre importantes relaciones causales. Lo sorprendente, lo contra intuitivo debería ser un criterio importante para las buenas hipótesis. El interés decae ante las numerosas replicaciones de una hipótesis; el interés se reaviva al buscar y encontrar factores contextuales que modifiquen el efecto. La importancia se relaciona con la frecuencia, gravedad y trascendencia social del problema implicado; los beneficios y aplicabilidad que se pueden anticipar de su resolución; o el avance del conocimiento en ese u otros campos y las futuras guías de investigación que probablemente se deriven. Por ejemplo, las hipótesis sobre el cáncer son más importantes que aquellas sobre varices porque son más las personas (y los fenómenos biológicos y psicológicos) que están afectadas más profundamente por el primero. Si bien, en ejemplos menos obvios, las ramificaciones de una teoría suelen ser difíciles de anticipar. Es fácil deducir la conexión de este criterio con los de concordancia, verosimilitud, independencia y generalización.

Calidad de las pruebas

Este criterio se refiere a la categorización de las pruebas científicas por la calidad de sus fuentes o tipos de estudios. Se relaciona con la validez interna y la fuerza “demostrativa” de causalidad derivada de los tipos de estudios o fuentes suministradoras de pruebas. Aunque considerado en parte en el criterio de concordancia (replicación con diferentes métodos o diseños), la acepción moderna del criterio de calidad de las pruebas introduce la clasificación de los tipos o diseños en orden descendente según su fortaleza, es decir, el grado en que estos consiguen el control de variables extrañas. Así, diversos grupos han establecido clasificaciones jerárquicas de la calidad de las pruebas científicas, fundamentalmente de cara a establecer pronósticos, priorizar intervenciones preventivas o evaluar tecnologías sanitarias, como el sistema del Centro de Medicina Basada en Pruebas de Oxford (anexo I).

La corroboración de una hipótesis causal es mayor cuanto más severas son las condiciones para contrastar; en este caso, la homogeneización o minimización de factores perturbadores (confusores) de la relación “causal” en el estudio contrastador, que puede conseguirse en mayor grado por la aleatorización en estudios de intervención o de aleatorización mendeliana; en un grado menor por restricción, equiparación, análisis estratificado o modelos matemáticos en estudios cuasi-experimentales, observacionales de cohortes o casos y controles; a continuación, los estudios poblacionales transversales (de correlación), y en un grado mínimo en estudios descriptivos, series de casos o experiencias clínicas.

El sistema GRADE de evaluación de las pruebas científicas establece cuatro niveles de calidad (alto, moderado, bajo y muy bajo) para el conjunto de las pruebas científicas, por lo que no se aplica a estudios aislados. (47) En primer lugar, establece un nivel de calidad de partida en función del diseño de los estudios: alta con estudios de intervención aleatorizados y baja con estudios observacionales. En segundo lugar, considera una serie de criterios cuya presencia reduce el nivel de calidad: riesgo de sesgo, falta de concordancia, relación indirecta, imprecisión y sesgo de publicación. Y en tercer lugar, considera una serie de criterios cuya presencia aumenta el nivel de calidad: magnitud del efecto, gradiente biológico y confusión residual posible que reduciría un efecto observado o sugeriría un efecto no observado.

En el campo de los estilos de vida (alimentación, actividad física, sueño, adicciones, relaciones sociales), los efectos estudiados son de pequeña magnitud (RR entre 0,8 y 1,2 habitualmente), el tiempo de latencia normalmente largo y la realización de estudios de intervención con frecuencia no es factible o ético. Por ello, se ha propuesto el uso de la llamada jerarquía de las pruebas científicas aplicadas a la medicina de estilos de vida. (48) HEALM (por sus siglas en inglés: *Hierarchies of evidence applied to lifestyle Medicine*) establece tres niveles de calidad: fuerte o decisiva, moderada o sugestiva, e insuficiente o no concluyente. Este sistema se plantea en primer lugar si la hipótesis causal es o no susceptible, por cuestiones éticas o de factibilidad, de ser contrastada con estudios de intervención. Si la respuesta es afirmativa, se recurre al sistema GRADE en caso de que se hayan llevado a cabo tales ensayos o al sistema del Centro de Medicina Basada en Pruebas de Oxford en caso contrario. Cuando no es posible contrastar la hipótesis directamente con estudios de intervención, se establecen los siguientes criterios: verosimilitud biológica, existencia de estudios de intervención que proporcionen pruebas indirectas de causalidad, existencia de estudios observacionales de alta calidad y gran tamaño muestral, que permitan generalizar sus resultados a grandes poblaciones, y existencia de estudios observacionales de alta calidad y períodos de seguimiento amplios, que permitan detectar efectos a largo plazo.

5. Práctica de la inferencia causal

Weed examina la práctica de la inferencia causal en epidemiología, y encuentra un uso relativamente pobre y heterogéneo de los criterios de causalidad de Hill, por lo que propugna un mayor uso de estos criterios, estudiar sus relaciones y desarrollar otros. (49) Dado un determinado conjunto de criterios, no está claro

cómo combinarlos o priorizarlos, ni si esto es realmente necesario o factible. Autores como Trichopoulos y MacMahon otorgan un papel esencial a la verosimilitud biológica ante efectos discordantes y generalmente débiles, y apuestan por la causalidad ante asociaciones fuertes y concordantes, aun en ausencia de apoyo biológico.⁽²⁾ Estos mismos autores apuntan que ante resultados no concluyentes, moderadamente compatibles con la hipótesis, puede ajustarse la principal medida de efecto no solo por confusores claros, sino también por variables con cuestionable (incluso improbable) potencial de confusión. Esto mejora el control de la confusión y la consiguiente ampliación del intervalo de confianza reduce la tentación de sobreinterpretación. Sin embargo, en ausencia de un conocimiento profundo de la estructura causal del desenlace de interés, no está claro cuál es la mejor estrategia de selección de las variables de ajuste para el control de la confusión, por lo que es recomendable llevar a cabo análisis de sensibilidad en busca de resultados compatibles que proporcionen mayor verosimilitud a la hipótesis causal.⁽⁴⁾ VanderWeele ha propuesto recientemente una serie de principios para la selección de variables de confusión, como base para la discusión y el futuro refinamiento de los mismos. ⁽⁵⁰⁾

En la práctica de la epidemiología y la salud pública también se empiezan a considerar conjuntamente los criterios de Hill y los de calidad de los estudios, es decir, a ponderar los criterios de causalidad a la luz de los diseños de las fuentes de datos, considerando adicional y específicamente los sesgos y los factores de confusión. Algunos autores sitúan en un primer escalón los estudios de mayor peso de diseño (estudios de intervención, estudios de series temporales, de aleatorización mendeliana y de cohortes), en segundo nivel los RR grandes y concordantes, y a los estudios de casos y controles, y en el último lugar los RR pequeños, aunque verosímiles, y los estudios transversales. Sin embargo, prácticamente siempre el juicio causal se realiza valorando pruebas de muchos y diferentes tipos de estudios. Así, se han demostrado los efectos del consumo de cigarrillos mediante experimentos con animales, grupos controlados y encuestas con seres humanos. De hecho, la investigación en humanos muchas veces comienza con pistas etiológicas de observaciones clínicas o estudios descriptivos, prosigue con estudios observacionales analíticos y, si estos parecen corroborar los hallazgos, se realiza una evaluación experimental, cuando es posible (agentes potencialmente beneficiosos). Por ejemplo, en 2010 se recomendaba iniciar el tratamiento de la infección por VIH en personas sintomáticas con un recuento de linfocitos CD4 inferior a 350 por milímetro cúbico de sangre. Una posible estrategia para reducir la transmisión del VIH consistía en iniciar la terapia antiretroviral inmediatamente después de obtener un test positivo a VIH, ya que reducía la concentración viral, pero no se recomendaba con recuento de linfocitos CD4 igual o superior a 500 por dudas en cuanto a la toxicidad de los fármacos y posible acumulación de resistencias, hasta que los análisis de varias

cohortes de VIH encontraron que la iniciación precoz del tratamiento resultaba en un menor riesgo de SIDA y muerte, resultados corroborados posteriormente en un ensayo clínico.⁽⁴⁾ Esta es una secuencia inversa a la evaluación de calidad de las pruebas científicas de las escalas jerárquicas. La consideración conjunta de los tres grandes tipos de estudios disponibles, la variedad de diseños (no la jerarquización y prima excesiva de los experimentos) es, como vimos en el criterio de concordancia, una regla que refuerza la inferencia. Como señala Weed, un problema adicional es que la selección, clasificación, definición e interpretación de criterios varían de usuario a usuario y de grupo a grupo, lo que puede producir juicios muy diferentes sobre la causación a partir de las mismas pruebas. ⁽⁴⁹⁾ De hecho, se suele decir que la valoración causal es materia de juicio subjetivo. De juicio, sí; subjetivo, no tanto. A pesar de la aparentemente subjetiva disparidad, el proceso inferencial se objetiva al abrirse la discusión a los demás. Las razones en pro y en contra se pueden sopesar y discutir; así se posibilita el acuerdo. El refinamiento de los criterios y su sistematización, su mayor reproducibilidad por todos, el mayor consenso sobre su definición y reglas de interpretación, mejorará la toma de decisiones causales, pero esta es siempre provisional, revisable: todos podemos errar en la misma dirección, o las pruebas pueden cambiar. Dada la cantidad de criterios existentes y la diversidad de su manejo, en la tabla 2 se indican los criterios más importantes, en un orden razonable según diversas perspectivas. Se prioriza el descarte razonable de explicaciones o factores no causales, y se distinguen los criterios que afectan a la validez interna, a la validez externa y a la comparación con otras pruebas. En general, los criterios más fuertes son la precedencia temporal, la magnitud razonable de la asociación (y su gradiente si se busca y encuentra), la concordancia con otros hallazgos y la verosimilitud biológica de la hipótesis.

Tabla 2. Cuestiones más relevantes para valorar la causalidad

A. Validez interna-Explicaciones no causales (o incluso ausencia de asociación)
1. Sesgos
2. Factores de confusión
3. Azar
B. Validez interna - Características positivas
4. Relación temporal
5. Fuerza
6. Dosis-respuesta
7. Concordancia
C. Validez externa - Generalizabilidad
8. Aplicación a otras poblaciones
D. Comparación con otras pruebas
9. Calidad de las pruebas
10. Concordancia
11. Verosimilitud

6. Referencias

1. Porta M (2014). A Dictionary of Epidemiology, 6th ed. New York: Oxford University Press.
2. MacMahon B, Trichopoulos D (1996). Epidemiology. Principles and methods. 2.^a ed. Boston: Little Brown.
3. Rothman KJ, Greenland S, Poole C, Lash TL (2008). Causation and causal inference. En: Rothman KJ, Greenland S, Lash TL, eds. Modern epidemiology. 3.^a ed. Filadelfia: Lippincott-Williams&Wilkins.

4. Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
5. Cookson W (1996). Cazadores de genes. La aventura del genoma. Madrid: Pirámide.
6. De Kruif P (2006). Los cazadores de microbios. 12.^a ed. México DF: Porrúa.
7. Glass DJ, Hall N. A brief history of the hypothesis. Cell. 2008;134:378-81.
8. Banegas JR, Rodríguez Artalejo F, Rey Calero J. Popper y el problema de la inducción en epidemiología. Rev Esp Salud Pública. 2000;74:327-39.
9. Popper KR (2002). The logic of scientific discovery. London: Routledge.
10. Bunge M. La investigación científica. Su estrategia y su filosofía. Barcelona: Ariel;1985.
11. Howson C, Urbach P. Bayesian reasoning in science. Nature 1991;350:373-4.
12. Parascandola M, Weed D L. Causation in epidemiology. J Epidemiol Community Health. 2001;55:905-2.
13. Rothman KJ (1988). Causal inference. Chestnut Hill: Epidemiology Resources Inc.
14. Giere RM. Understanding scientific reasoning. 5.^a edición. Belmont: Thomson/Wadsworth;2006.
15. Elwood JM. Causal relationships in Medicine. Nueva York: Oxford University Press;1992.
16. Greenland S, editor. Evolution of epidemiologic ideas. Chestnut Hill: Epidemiology Resources Inc.; 1987.
17. Hernán MA. Invited commentary: hypothetical interventions to define causal effects-afterthought or prerequisite? Am J Epidemiol. 2005;162:618-20.
18. Rubin D. Statistics and causal inference: comment: which ifs have causal answers. J Am Stat Assoc. 1986;81:961-2.
19. Schwartz S, Prins SJ, Campbell UB, Gatto NM. Is the “well-defined intervention assumption” politically conservative? Soc Sci Med. 2016;166:254-7.
20. Evans S. Causation and disease: the Henle-Koch postulates revisited. Yale J Biol Med. 1976;49:175-95.
21. Szklo M, Nieto FJ (2019). Epidemiology: beyond the basics. Fourth edition. Burlington, Massachusetts : Jones & Bartlett Learning.
22. MacMahon B, Pugh TF, Ipsen J (1960). Epidemiology: Principles and Methods. Boston, MA: Little, Brown and Company.

23. Doll R, Peto R. The causes of cancer. Oxford: Oxford University Press;1981.
24. Hill AB. The environment and disease: association or causation? Proc Roy Soc Med. 1965;58:295-300.
25. Schwartz J. Air pollution and daily mortality: a review and meta analysis. Environ Res. 1994;64:36-52.
26. Remington PL, Brownson RC, Wegner MV, eds. Chronic disease epidemiology and control. 3.^a ed. Washington: American Public Health Association;2010.
27. Potischman N, Weed DL. Causal criteria in nutritional epidemiology. Am J Clin Nutr. 1999;69:S1309-14.
28. Kassirer JP, Angell M. Risk adjustment or risk avoidance? N Engl J Med. 1998;339:1925-6.
29. Ioannidis JPA. Exposure-wide epidemiology: revisiting Bradford Hill. Stat Med. 2016;35:1749-62.
30. Buckley DI, Fu R, Freeman M, Rogers K, Helfand M. C-reactive protein as a risk factor for coronary heart disease: a systematic review and meta-analyses for the U.S. Preventive Services Task Force. Ann Intern Med. 2009;151:483-95.
31. ISIS-4 (Fourth International Study of Infarct Survival Collaborative Group). ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. Lancet. 1995;345:669-85.
32. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg Ch, Stefanick ML, *et al*; Women's Health Initiative. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA. 2002;288:321-33.
33. Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N Engl J Med. 1994;330:1029-35.
34. He J, Vupputuri S, Allen K, Prerost MR, Hughes J, Whelton PK. Passive smoking and the risk of coronary heart disease--a meta-analysis of epidemiologic studies. N Engl J Med. 1999;340:920-6.
35. Barker DJ. The origins of the developmental origins theory . J Intern Med. 2007;261:412-7.
36. Dawber TR. The Framingham Study. The epidemiology of atherosclerosis disease. Cambridge: Harvard University Press; 1980.

37. Loeb LA, Ernster VL, Warner KE, Abbotts J, Laszlo J. Smoking and lung cancer: an overview. *Cancer Res.* 1984;44:5940-58.
38. Xavier Bosch F, de Sanjosé S. The epidemiology of human papillomavirus infection and cervical cancer. *Dis Markers.* 2007;23:213-27.
39. Ross R. Atherosclerosis: an inflammatory disease. *N Engl J Med.* 1999;340:115-26.
40. Cerda J, Valdivia G. John Snow, la epidemia de cólera y el nacimiento de la epidemiología moderna. *Rev Chil Infect.* 2007;24:331-4.
41. Susser M. What is a cause and how do we know one? A grammar for pragmatic epidemiology. *Am J Epidemiology.* 1991;133:635-48.
42. Jérolon A, Baglietto L, Birmelé E, Alarcon F, Perduca V. Causal mediation analysis in presence of multiple mediators uncausally related. *Int J Biostat.* 2020;17(2):191-221.
43. Song Y, Zhou X, Kang J, *et al.* Bayesian hierarchical models for high-dimensional mediation analysis with coordinated selection of correlated mediators. *Stat Med.* 2021;40:6038-56.
44. Chlebowski RT, Aragaki AK, Anderson GL, Pan K, Neuhouser ML, Manson JE, *et al.*; Women's Health Initiative. Dietary Modification and Breast Cancer Mortality: Long-Term Follow-Up of the Women's Health Initiative Randomized Trial. *J Clin Oncol.* 2020;38:1419-28.
45. The Pooling Project Research Group. Relationship of blood pressure, serum cholesterol, smoking habits, relative weight, and ECG abnormalities to incidence of major coronary events: Final report of the pooling Project. *J Chronic Dis.* 1978; 31:201-306.
46. Ross R. The pathogenesis of atherosclerosis: a perspective for the 1990s. *Nature.* 1993;362:801.
47. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, *et al.* GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol.* 2011;64:401-6.
48. Katz DL, Karlsen MC, Chung M, Shams-White MM, Green LW, Fielding J, *et al.* Hierarchies of evidence applied to lifestyle Medicine (HEALM): introduction of a strength-of-evidence approach based on a methodological systematic review. *BMC Med Res Methodol.* 2019;19:178.
49. Weed DL. On the use of causal criteria. *Int J Epidemiol.* 1997;26:1137-41.
50. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol.* 2019;34(3):211-219.

ANEXO I. Centro de Oxford para la medicina basada en pruebas: niveles de prueba

Nivel	Terapia / Prevención, Etiología / Daño	Pronóstico	Diagnóstico	Diagnóstico diferencial / Estudio de prevalencia de síntomas	Análisis económicos y de decisión
1a	Revisión sistemática (con homogeneidad)* de ensayos clínicos.	Revisión sistemática (con homogeneidad)* de estudios de cohorte de inicio; CDR" validado en diferentes poblaciones.	Revisión sistemática (con homogeneidad)* de estudios diagnósticos de nivel 1; CDR" con estudios 1b de diferentes centros clínicos.	Revisión sistemática (con homogeneidad)* de estudios prospectivos de cohortes.	Revisión sistemática (con homogeneidad)* de estudios económicos de nivel 1.
1b	Ensayo clínico aleatorizado (ECA) individual (con intervalo de confianza estrecho).	Estudio de cohorte de casos incidentes con > 80% de seguimiento; RDC" validado en una sola población.	Estudio de cohorte de validación** con buenos estándares de referencia" " " ; o RDC" probada en un centro clínico.	Estudio de cohorte prospectivo con buen seguimiento.***	Análisis basado en costos o alternativas clínicamente sensibles; revisión(es) sistemática(s) de las pruebas, incluyendo análisis de sensibilidad de múltiples vías.
1c	Todo o nada.§	Serie de casos todos o ninguno.	SpPins y SnNouts absolutos" " "	Serie de casos todos o ninguno.	Análisis absolutos de mejor o peor valor " " " "
2a	Revisión sistemática (con homogeneidad)* de estudios de cohortes.	Revisión sistemática (con homogeneidad)* de estudios de cohortes retrospectivos o grupos control no tratados en ECA.	Revisión sistemática (con homogeneidad)* de estudios de diagnóstico de nivel >2.	Revisión sistemática (con homogeneidad)* de 2b y mejores estudios.	Revisión sistemática (con homogeneidad)* de estudios económicos de nivel > 2.
2b	Estudio de cohorte individual (incluyendo ECA de baja calidad; p. ej., <80% de seguimiento)	Estudio de cohorte retrospectivo o seguimiento de pacientes control no tratados en un ECA; Derivación de RDC" o validada solo en muestra dividida§§§	Estudio de cohorte exploratorio** con buenos estándares de referencia" " " ; RDC" después de la derivación o validado solo en bases de datos muestras divididas.§§§	Estudio de cohorte retrospectivo o seguimiento deficiente.	Análisis basado en costos o alternativas clínicamente sensibles; revisión(es) limitada(s) de las pruebas o estudios únicos; incluyendo análisis de sensibilidad de múltiples vías.
2c	Investigación de resultados; estudios ecológicos.	Investigación de resultados.		Estudios ecológicos.	Auditoría o investigación de resultados.

Nivel	Terapia / Prevención, Etiología / Daño	Pronóstico	Diagnóstico	Diagnóstico diferencial / Estudio de prevalencia de síntomas	Análisis económicos y de decisión
3a	Revisión sistemática (con homogeneidad)* de estudios de casos y controles.		Revisión sistemática (con homogeneidad)* de 3b y mejores estudios.	Revisión sistemática (con homogeneidad)* de 3b y mejores estudios.	Revisión sistemática (con homogeneidad)* de 3b y mejores estudios.
3b	Estudio de casos y controles.		Estudio no consecutivo o sin estándares de referencia aplicados consistentemente.	Estudio de cohortes no consecutivos o población muy limitada.	Análisis basado en alternativas o costos limitados, estimaciones de datos de baja calidad, pero que incluyen análisis de sensibilidad que incorporan variaciones clínicamente sensibles.
4	Serie de casos (y estudios de cohortes y de casos y controles de baja calidad). ^{§§}	Serie de casos (y estudios de cohortes de pronóstico de baja calidad). ^{***}	Estudio de casos y controles, estándar de referencia deficiente o no independiente.	Serie de casos o estándares de referencia reemplazados.	Análisis sin análisis de sensibilidad.
5	Opinión de expertos sin evaluación crítica explícita, o basada en fisiología, investigación de banco o primeros principios.	Opinión de expertos sin evaluación crítica explícita, o basada en fisiología, investigación de banco o primeros principios.	Opinión de expertos sin evaluación crítica explícita, o basada en fisiología, investigación de banco o primeros principios.	Opinión de expertos sin evaluación crítica explícita, o basada en fisiología, investigación de banco o primeros principios.	Opinión de expertos sin valoración crítica explícita, o basada en teoría económica o primeros principios.

Produced by Bob Phillips, Chris Ball, Dave Sackett, Doug Badenoch, Sharon Straus, Brian Haynes, Martin Dawes since November 1998. Updated by Jeremy Howick, March 2009. <https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009>

Notas

Los usuarios pueden agregar un signo menos (-) al nivel asignado en los siguientes casos:

- un único resultado con un intervalo de confianza amplio.
- una revisión sistemática con heterogeneidad problemática.

Dicha evidencia no es concluyente y, por lo tanto, solo puede generar recomendaciones de grado D.

Bueno, mejor, malo y peor se refieren a las comparaciones entre tratamientos en términos de sus riesgos y beneficios clínicos.

* Por homogeneidad, nos referimos a una revisión sistemática que está libre de variaciones preocupantes (heterogeneidad) en las direcciones y grados de resultados entre estudios individuales. No todas las revisiones sistemáticas con heterogeneidad estadísticamente significativa deben ser preocupantes y no todas las heterogeneidades preocupantes deben ser estadísticamente significativas. Los estudios que muestran una heterogeneidad preocupante deben etiquetarse con un signo menos al nivel asignado.

"Regla de Decisión Clínica (algoritmos o sistemas de puntuación que conducen a una estimación de pronóstico o a una categoría de diagnóstico). Consulte la nota anterior para obtener consejos sobre cómo comprender, calificar y utilizar ensayos u otros estudios con amplios intervalos de confianza.

§ Todos los pacientes murieron antes de que el Rx estuviera disponible, pero algunos ahora sobreviven con él; o algunos pacientes fallecieron antes de que el Rx estuviera disponible, pero ahora ninguno muere con él.

§§ Por estudio de cohorte de mala calidad nos referimos a uno que no pudo definir claramente los grupos de comparación o no pudo medir las exposiciones y los resultados de la misma manera objetiva (preferiblemente ciega) en expuestos y no expuestos o no pudo identificar o controlar adecuadamente los factores de confusión conocidos o no realizó un seguimiento suficientemente largo y completo de los pacientes. Por estudio de casos y controles de mala calidad nos referimos a uno que no pudo definir claramente los grupos de comparación o no pudo medir las exposiciones y los resultados de la misma manera objetiva (preferiblemente ciega) en casos y controles o no pudo identificar o controlar adecuadamente factores de confusión conocidos.

§§§ La validación de muestra dividida se logra recopilando toda la información en un solo tramo y luego dividiéndola artificialmente en muestras de derivación y validación.

" " Un SpPin absoluto es un hallazgo diagnóstico cuya especificidad es tan alta que un resultado positivo determina el diagnóstico. Un SnNout absoluto es un hallazgo diagnóstico cuya sensibilidad es tan alta que un resultado negativo descarta el diagnóstico.

" " " Los buenos estándares de referencia son independientes de la prueba y se aplican a ciegas u objetivamente a todos los pacientes. Los estándares de referencia deficientes se aplican al azar, pero siguen siendo independientes de la prueba. El uso de un estándar de referencia no independiente (donde la "prueba" está incluida en la "referencia", o donde la "prueba" afecta a la "referencia") implica un estudio de nivel 4.

" " " " Los tratamientos de mejor valor son claramente igual de buenos pero más baratos, o mejores al mismo costo o a un costo reducido. Los tratamientos de peor valor son tan buenos y más caros, o peores e igual o más caros.

* * Los estudios de validación prueban la calidad de una prueba diagnóstica específica, con base en evidencia previa. Un estudio exploratorio recopila información y rastrea los datos (por ejemplo, utilizando un análisis de regresión) para encontrar qué factores son estadísticamente significativos.

*** Por estudio de cohorte de pronóstico de mala calidad nos referimos a uno en el que el muestreo estuvo sesgado a favor de los pacientes que ya tenían el resultado objetivo o la medición de los resultados se logró en < 80 % de los pacientes del estudio o los resultados se determinaron de forma no cegada, de manera no objetiva o sin corrección por factores de confusión.

**** Un buen seguimiento en un estudio de diagnóstico diferencial es > 80 %, con tiempo adecuado para que surjan diagnósticos alternativos (por ejemplo, 1 a 6 meses para cuadros agudos y 1 a 5 años para cuadros crónicos).

Grados de recomendación

- A. Estudios consistentes de nivel 1.
- B. Estudios consistentes de nivel 2 o 3 o extrapolaciones de estudios de nivel 1.
- C. Estudios de nivel C 4 o extrapolaciones de estudios de nivel 2 o 3.
- D. Pruebas de nivel D 5 o estudios preocupantemente inconsistentes o no concluyentes de cualquier nivel.

Las extrapolaciones se dan cuando los datos se usan en situaciones que tienen diferencias potencialmente importantes desde el punto de vista clínico con el estudio original.

Glosario de términos básicos en epidemiología

JAVIER DAMIÁN, MIGUEL ÁNGEL ROYO BORDONADA, FERNANDO JOSÉ GARCÍA LÓPEZ

Este glosario incluye una sucinta selección de términos para facilitar la consolidación de algunos conceptos importantes. Para una consulta más exhaustiva de los términos utilizados en epidemiología se recomienda el diccionario de epidemiología¹. Agradecemos a los profesores que participan en este volumen y otros colegas en el Centro Nacional de Epidemiología sus comentarios y sugerencias. Las definiciones no pretenden aportar explicaciones detalladas, ya que esto queda para los textos, sino afianzar elementos esenciales de los conceptos. Asimismo, suscribimos las palabras de Neil Pearce —un reconocido epidemiólogo— escritas en el preámbulo del diccionario de epidemiología¹:

«Escribir un diccionario en cualquier disciplina científica es una empresa arriesgada, porque los científicos suelen discrepar. La naturaleza de la ciencia no es llegar a un consenso, sino hacer avanzar nuestros conocimientos sometiendo las ideas contradictorias a un examen crítico. Lo mismo ocurre con la definición de los conceptos que utilizamos. Ningún diccionario puede satisfacer a todo el mundo, ni debe intentarlo».

- **ALEATORIZACIÓN.** Procedimiento por el que se asigna a los sujetos de estudio a los grupos experimentales mediante el azar. Técnica habitual en ensayos clínicos y otros diseños experimentales orientada a minimizar la confusión.
- **ANÁLISIS ESTRATIFICADO.** Estimación separada de la asociación o efecto en determinados subgrupos, con objeto de controlar la confusión o identificar la modificación del efecto por una tercera variable.
- **ASOCIACIÓN ESTADÍSTICA.** Dos variables están asociadas si el conocimiento del valor de una aporta alguna información sobre el valor de la otra. La asociación es causal si un cambio en una variable ocasiona cambios en la otra.
- **CAUSA.** Suceso, condición o característica (antecedente) que produce un efecto o aumenta la probabilidad de que se produzca un efecto.
- **COHORTE.** Estrictamente es un conjunto de personas que comparten una característica al comienzo de un seguimiento. Se utiliza también para referirse a poblaciones dinámicas bien definidas que son seguidas en el tiempo para registrar la aparición de casos de enfermedad.

¹ Porta M, Greenland S, Hernán M, dos Santos Silva I, Last JM, editores. A dictionary of epidemiology. Sexta edición. Oxford: Oxford University Press; 2014.

- **CONFUSIÓN.** Distorsión en la medida de efecto originada por la acción de una tercera variable o conjunto de variables. En general hay confusión cuando el parámetro de desenlace (por ejemplo, el riesgo medio de enfermar) obtenido en personas no expuestas es diferente al que tendrían las personas expuestas si no estuvieran expuestas. En un gráfico con la estructura causal del problema codificada, la confusión se debe a la existencia de un camino abierto por la puerta de atrás entre la exposición y el desenlace.
- **CRIBADO.** Intervención de salud pública cuyo objetivo es la detección precoz de una enfermedad.
- **DETERMINANTE.** Característica de los sujetos de estudio de la cual depende algún parámetro de frecuencia de enfermedad, como una tasa de incidencia. Es un término sinónimo de causa derivado de los modelos deterministas de causalidad.
- **DIFERENCIA DE RIESGOS.** Diferencia entre el riesgo de enfermar en las personas expuestas y no expuestas, o expuestas en distinto grado, a un factor de riesgo. Se ha denominado clásicamente como riesgo atribuible.
- **EPIDEMIOLOGÍA.** Disciplina científica que se ocupa de la medición de aspectos relacionados con la salud en poblaciones específicas, lo que incluye el estudio de los determinantes que influyen en esos procesos, y la aplicación de este conocimiento para el control de problemas de salud relevantes.
- **EFFECTO ABSOLUTO (DE UNA EXPOSICIÓN SOBRE UNA ENFERMEDAD).** Diferencia entre la incidencia –riesgo o tasa– en población expuesta y la incidencia que tendría esta población si no estuviera expuesta. Representa la cantidad absoluta del riesgo o tasa en población expuesta que es atribuible a la exposición.
- **EFFECTO RELATIVO (DE UNA EXPOSICIÓN SOBRE UNA ENFERMEDAD).** Efecto absoluto dividido entre la incidencia basal (la incidencia que tendría la población expuesta si no estuviera expuesta). Representa el exceso de la razón de riesgos (o de la razón de tasas) sobre 1, el valor nulo de la medida.
- **EMPAREJAMIENTO.** Técnica de diseño en la que se seleccionan parejas que tienen las mismas categorías de determinados factores de confusión con el objeto de hacer más eficiente el control de los mismos en algunas situaciones. Se utiliza fundamentalmente en estudios de casos y controles, aunque también en estudios de cohortes. En los primeros es necesario realizar un análisis específico que corrija el sesgo que produce esta forma de selección.

- **ENFERMEDAD.** Alteración de la salud asociada a un conjunto definido de signos y síntomas. Término usado aquí de forma genérica para referirse a la variable dependiente, desenlace o resultado principal de un estudio.
- **ENSAYO CLÍNICO.** Estudio experimental realizado en personas con determinada enfermedad o dolencia para valorar la eficacia (o efectividad) y la seguridad de una intervención.
- **ENSAYO COMUNITARIO.** Ensayo por conglomerados en el que las unidades de aleatorización son comunidades enteras de gran tamaño, como barrios o ciudades.
- **ENSAYO CRUZADO.** Estudio experimental en el que cada sujeto recibe las dos intervenciones a comparar, la experimental y la de control, en distintos periodos determinados aleatoriamente.
- **ENSAYO DE CAMPO.** Estudio realizado en personas libres de una enfermedad o dolencia, dirigido a valorar la eficacia (o efectividad) y seguridad de intervenciones experimentales preventivas.
- **ERROR ALEATORIO.** Desviación producida en la estimación de un parámetro con respecto a su verdadero valor debida únicamente a variabilidad aleatoria intrínseca al muestreo o a los errores aleatorios de medición. En general se puede calcular su magnitud pero no su sentido.
- **ERROR SISTEMÁTICO.** Sinónimo de sesgo. Es la desviación producida en la estimación de un parámetro con respecto a su verdadero valor debida a factores que permiten estimar su sentido, aunque en general no se puede estimar su magnitud.
- **ESPECIFICIDAD DE UNA PRUEBA DIAGNÓSTICA.** Probabilidad de que un no enfermo tenga un resultado negativo al aplicarle la prueba.
- **ESTANDARIZACIÓN.** Análisis estratificado en el que se usan distribuciones estándar de factores de confusión que pueden ser diferentes de las distribuciones de los datos del estudio.
- **ESTUDIO ANALÍTICO.** Estudio cuyo objetivo es comprobar una hipótesis.
- **ESTUDIO CASO-COHORTE.** Estudio que se realiza sobre una cohorte establecida, seleccionando los casos que se producen en el seguimiento y una muestra de sujetos de la cohorte al inicio del seguimiento.
- **ESTUDIO CUASI-EXPERIMENTAL.** Estudio para evaluar el efecto de una intervención sin grupo control (experimentos no controlados) o donde la intervención ha sido asignada mediante un mecanismo distinto al azar (experimentos controlados sin aleatorizar). Cuando la intervención se ha

producido de forma espontánea hablamos de experimentos naturales. Paradigma de los diseños para valorar cambios en políticas de salud pública.

- **ESTUDIO DE CASOS CRUZADOS.** Estudio realizado solo en casos de enfermedad para medir efectos de exposiciones cambiantes en el tiempo y elevaciones transitorias del riesgo. Se miden los niveles de exposición en periodos de riesgo especificados previos a la aparición de la enfermedad (periodos caso) y se comparan con los niveles de exposición habituales del sujeto (periodos control).
- **ESTUDIO DE CASOS Y CONTROLES ANIDADO EN UNA COHORTE.** Estudio en el seno de una cohorte establecida en la que se seleccionan los casos y una muestra aleatoria de controles entre los sujetos que están a riesgo enfermar cuando ocurre cada caso. La selección de controles de esta forma es proporcional a la cantidad persona-tiempo de exposición en la cohorte, por lo que la *odds ratio* estima la razón de tasas de incidencia.
- **ESTUDIO DE CASOS Y CONTROLES.** Estudio epidemiológico en el que el criterio de selección de los sujetos de estudio es presentar o no la enfermedad o desenlace.
- **ESTUDIO DE COHORTES.** Estudio epidemiológico en el que el criterio de selección de los sujetos de estudio es presentar o no la exposición de interés.
- **ESTUDIO DESCRIPTIVO.** Estudio cuyo objetivo es describir la situación para facilitar la toma de decisiones o para facilitar el planteamiento de hipótesis. Sin ánimo de comprobar hipótesis etiológicas.
- **ESTUDIO CUANTITATIVO.** Estudio en el que el conocimiento se obtiene mediante el análisis e interpretación de datos numéricos.
- **ESTUDIO CUALITATIVO.** Estudio descriptivo en el que el conocimiento se obtiene mediante el análisis e interpretación de información no numérica que permite elaborar un discurso, argumentación o constructo.
- **ESTUDIO ECOLÓGICO.** Aquel en el que las unidades de análisis son agrupaciones de sujetos.
- **ESTUDIO EXPERIMENTAL.** Estudio de cohortes en el que el criterio de selección es haber recibido una determinada intervención experimental.
- **ESTUDIOS NO EXPERIMENTALES.** Todos los diseños que no estudian intervenciones experimentales. Tradicionalmente denominados estudios observacionales.
- **ESTUDIOS PROSPECTIVOS/RETROSPECTIVOS.** Denominación confusa y controvertida. Es prospectivo cuando se recoge la información

sobre exposiciones y otras variables después del diseño del estudio y se completa con el paso del tiempo, recogiendo información sobre desenlaces y cambios en la exposición. Es retrospectivo si la recogida de información sobre exposiciones, covariables y desenlace ha ocurrido antes del diseño del estudio. Otra definición considera prospectivo un estudio en el que la medición o registro de la variable de exposición es anterior a la ocurrencia de enfermedad y retrospectivo si se recoge información sobre la exposición cuando ya ha ocurrido la enfermedad. En ambos casos, prospectivo confiere una potencial ventaja de calidad (en cuanto a validez) *a priori*.

- **ESTUDIO TRANSVERSAL.** Estudio en el que la selección de sujetos y la determinación de sus niveles de exposición y estatus de enfermedad se realizan virtualmente en el mismo momento. También se han llamado estudios de prevalencia.
- **EXPERIMENTO.** Modificación protocolizada de la realidad dirigida a ganar conocimiento.
- **EXPOSICIÓN.** Característica externa o ambiental que afecta al individuo y que es un posible determinante de la enfermedad o desenlace de interés. Se suele usar de forma genérica para referirse a cualquier tipo de determinante —variable independiente, hábito, característica personal, etc.— que es objeto de estudio.
- **FACTOR DE CONFUSIÓN (DE LA ASOCIACIÓN ENTRE UNA EXPOSICIÓN Y UN DESENLACE).** Variable asociada con la exposición en la población y con el desenlace en la población no expuesta, que no es un mediador en la cadena causal entre exposición y desenlace. Estas condiciones son necesarias para considerar una variable como factor de confusión, pero no suficientes. En un gráfico con la estructura causal del problema codificada, los factores de confusión son las causas comunes de la exposición y el desenlace.
- **FACTOR DE COLISIÓN.** En un gráfico con la estructura causal del problema codificada, un factor de colisión es un efecto de dos o más variables, que se representa como un nodo (variable) en un camino entre la exposición y el desenlace al que apuntan dos o más flechas que colisionan entre sí.
- **FACTOR DE RIESGO.** Característica de los sujetos de estudio de la cual depende algún parámetro de frecuencia de enfermedad, como una tasa de incidencia. Es un término sinónimo de causa derivado de los modelos probabilísticos de causalidad.
- **FRACCIÓN ETIOLÓGICA.** Fracción de casos de enfermedad para los cuales la exposición desempeña un papel causal en su desarrollo. No puede

estimarse solo con datos epidemiológicos ya que depende además del modelo biológico de desarrollo de la enfermedad que se asuma como válido. Si se utilizan riesgos la proporción atribuible en expuestos se puede considerar un límite inferior de la fracción etiológica en los expuestos.

- **GENERALIZABILIDAD.** Grado en que los resultados de un estudio se pueden aplicar a poblaciones distintas a la estudiada. También denominada VALIDEZ EXTERNA.
- **INCIDENCIA ACUMULADA.** Proporción de casos nuevos de enfermedad u otro desenlace que se producen en un periodo de tiempo definido entre los sujetos de estudio al inicio del periodo.
- **INTERACCIÓN (DE DOS VARIABLES).** Efecto conjunto de dos exposiciones que es diferente del esperado por la combinación de sus efectos individuales. Dos exposiciones interaccionan cuando el efecto de una de ellas en la enfermedad depende de algún modo de la presencia (o ausencia) de la otra.
- **MEDIDAS DE ASOCIACIÓN.** Diferencias o razones entre las medidas de frecuencia de enfermedad de dos o más poblaciones.
- **MEDIDAS DE EFECTO.** Diferencias o razones entre las medidas de frecuencia de enfermedad de la misma población bajo diferentes supuestos para evaluar asociaciones causales. Son medidas teóricas ya que uno de los supuestos es contrafáctico y por tanto no puede ser observado. En la práctica, se obtienen comparando distintas poblaciones en condiciones de intercambiabilidad.
- **METANÁLISIS.** Técnica estadística utilizada en los estudios de compilación cuyo objetivo es la estimación de un parámetro—por ejemplo, una razón de riesgos—a partir de los resultados de varios estudios que estiman el mismo parámetro.
- **MODIFICACIÓN DE (LA MEDIDA DE) EFECTO.** Cambio en el valor de la medida de efecto de la exposición en distintos subgrupos. Por ejemplo, una diferencia de riesgos (entre expuestos y no expuestos) distinta en adultos mayores y jóvenes o una razón de riesgos (entre expuestos y no expuestos) diferente en mujeres y hombres.
- **MUESTREO PROBABILÍSTICO.** Procedimiento de selección en el que se conoce la probabilidad de cada individuo del marco muestral de ser incluido en una muestra. Es el mejor procedimiento para obtener una muestra representativa.
- **ODDS.** Término del idioma inglés ampliamente usado en epidemiología para describir la razón entre la probabilidad de un suceso y la probabilidad

del suceso complementario. Indica por tanto cuánto más o menos probable es que ocurra un suceso a que no ocurra.

- **ODDS RATIO.** Razón de dos *odds*. Se usa como medida de asociación o efecto en los estudios de casos y controles. Según sea el procedimiento de muestreo de controles, puede servir como estimador de la razón de tasas o de la razón de riesgos en la cohorte subyacente.
- **PARÁMETRO.** Valor poblacional de una dimensión de interés, como una media, una razón de riesgos o una prevalencia, susceptible de ser estimado mediante estudios que normalmente utilizan muestras.
- **PERIODO DE INDUCCIÓN.** Tiempo entre el comienzo de la causa y la ocurrencia del efecto.
- **PERIODO DE LATENCIA.** Tiempo entre el comienzo de la enfermedad y su diagnóstico.
- **PERSONAS-TIEMPO.** Cantidad total de tiempo de observación aportado por las personas participantes en un estudio. Una misma cantidad puede obtenerse observando menos personas durante más tiempo o más personas durante menos tiempo, dentro de unos límites razonables.
- **POBLACIÓN ABIERTA.** También llamada dinámica. Grupo de personas en estudio que admite entradas y salidas de personas según características de pertenencia (persona, lugar o tiempo).
- **POBLACIÓN CERRADA O FIJA.** Grupo de personas en estudio que no admite nuevos miembros. Normalmente el grupo es una cohorte, es decir, un grupo de personas con características de pertenencia que no cambian.
- **PRECISIÓN (DE UNA MEDIDA DE ASOCIACIÓN O EFECTO).** Es una medida de la variabilidad aleatoria asociada a una estimación. Se valora mediante intervalos de confianza (en estadística frecuentista) o intervalos de credibilidad (en estadística bayesiana). Intervalos amplios denotan alta variabilidad aleatoria y, por tanto, poca precisión.
- **PREVALENCIA.** Proporción de personas de una población que tienen una enfermedad, o característica, en un momento definido. Si es estable, estima la probabilidad de seleccionar al azar a un sujeto de la población que tenga la característica en cuestión.
- **PROPORCIÓN.** Razón en la que el numerador es parte del denominador y, por tanto, expresa cuantitativamente la parte de un total.
- **PROPORCIÓN ATRIBUIBLE (A UN FACTOR DE RIESGO) EN EXPUESTOS.** Proporción del total de riesgo (o tasa) en expuestos atribuible a la exposición. Puede dar una idea sobre la probabilidad de que una per-

sona expuesta y enferma deba su enfermedad a la exposición, si bien esta probabilidad no puede estimarse solo con datos epidemiológicos, ya que depende del modelo biológico de desarrollo de la enfermedad que se asuma como válido.

- **PROPORCIÓN ATRIBUIBLE (A UN FACTOR DE RIESGO) POBLACIONAL.** Proporción del total de riesgo (o tasa) en una población de expuestos y no expuestos que es atribuible a la exposición.
- **PROPORCIÓN PREVENIDA POR LA EXPOSICIÓN.** Proporción de casos potenciales que se hubieran producido en ausencia de exposición y que fueron prevenidos por la exposición. Si es en un grupo exclusivo de expuestos, la proporción prevenida se suele denominar eficacia de la intervención, y si esta es una vacuna, se denomina eficacia vacunal.
- **RAZÓN.** Cociente de dos números.
- **RAZÓN DE MORTALIDAD (O MORBILIDAD) ESTANDARIZADA.** Razón de dos tasas (o dos riesgos) estandarizadas por la distribución de una de ellas, que es la población índice o problema.
- **RAZÓN DE RIESGOS.** Razón entre el riesgo de enfermar en las personas con una categoría de exposición de un factor de riesgo y el correspondiente en las personas con una categoría de referencia del mismo factor de riesgo. Se ha denominado clásicamente como riesgo relativo.
- **RAZÓN DE VEROSIMILITUD DE UNA PRUEBA DIAGNÓSTICA NEGATIVA.** Razón entre la probabilidad de un resultado negativo en presencia de enfermedad ($1 - \text{sensibilidad}$) y su probabilidad en ausencia de enfermedad (especificidad). Indica cuánto más (o menos) probable es un resultado negativo en presencia de enfermedad que en su ausencia.
- **RAZÓN DE VEROSIMILITUD DE UNA PRUEBA DIAGNÓSTICA POSITIVA.** Razón entre la probabilidad de un resultado positivo en presencia de enfermedad (sensibilidad) y su probabilidad en su ausencia ($1 - \text{especificidad}$). Indica cuánto más (o menos) probable es un resultado positivo en presencia de enfermedad que en su ausencia.
- **REVISIÓN SISTEMÁTICA.** Investigación científica en la que la unidad de análisis son los estudios individuales o primarios. Las revisiones sistemáticas tienen como objetivo reunir y sintetizar todo el conocimiento disponible sobre una pregunta claramente formulada, con métodos sistemáticos y explícitos para identificar, seleccionar y evaluar críticamente la investigación relevante, y para recopilar y analizar datos de los estudios primarios incluidos en la revisión.

- RIESGO. Probabilidad de enfermar. Una forma de estimar su valor medio en un grupo es mediante la incidencia acumulada.
- RIESGO ATRIBUIBLE. Denominación clásica de la diferencia de riesgos. Suele confundirse con la proporción atribuible, por lo que es mejor utilizar el término diferencia de riesgos.
- RIESGO RELATIVO. Término inapropiado que se utiliza indistintamente, de forma genérica, para referirse a la razón de dos tasas, de dos riesgos e incluso de dos *odds* (*odds ratio*). Es inapropiado porque no se trata de un riesgo (que es una probabilidad) y porque no distingue el tipo de medida de frecuencia que se compara.
- SENSIBILIDAD DE UNA PRUEBA DIAGNÓSTICA. Probabilidad de que un enfermo tenga una prueba positiva.
- SESGO. Error sistemático en los resultados de un estudio con respecto a los reales en la población objetivo. En presencia de sesgo, el valor esperado de un parámetro estimado por los datos empíricos no coincide con el verdadero valor del parámetro. A diferencia del error aleatorio el sesgo, afecta a los resultados en un sentido identificable.
- SESGO DE INFORMACIÓN. Error sistemático que se producen durante la recogida e interpretación de información de las variables de un estudio.
- SESGO DE SELECCIÓN. Error sistemático derivado del proceso de selección de los sujetos que serán objeto de estudio.
- TASA DE INCIDENCIA. Razón entre el número de casos nuevos en un periodo y la cantidad persona-tiempo de observación en que pudieron ocurrir. Expresa la velocidad de ocurrencia de la enfermedad en la población.
- TASA DE MORTALIDAD. Tasa de incidencia de fallecimientos.
- TASA DE LETALIDAD (DE UNA ENFERMEDAD). Riesgo de morir entre las personas enfermas. Aunque se sigue usando ese término, no es una tasa sino una incidencia acumulada. Por tanto, tiene que especificarse el periodo de riesgo para valorar su magnitud.
- TASA, RIESGO O PREVALENCIA ESTANDARIZADA. Tasa, riesgo o prevalencia que tendría una población si en lugar de tener su propia distribución de determinadas variables, potenciales factores de confusión, tuviera otra, llamada estándar. Son mediadas hipotéticas, solo útiles para comparar con otras medidas que hayan sido estandarizadas por la misma distribución.

- VALIDEZ (DE UN ESTUDIO EPIDEMIOLOGICO). Es el grado en el que un estudio está libre de sesgos y confusión.
- VALIDEZ EXTERNA. Grado en que los resultados de un estudio se pueden aplicar a poblaciones distintas a la estudiada. También denominada GENERALIZABILIDAD.
- VALOR PREDICTIVO DE UNA PRUEBA DIAGNÓSTICA NEGATIVA. Probabilidad de que una persona con una prueba negativa no tenga la enfermedad.
- VALOR PREDICTIVO DE UNA PRUEBA DIAGNÓSTICA POSITIVA. Probabilidad de que una persona con una prueba positiva tenga la enfermedad.
- VALOR NULO (DE UN PARÁMETRO O MEDIDA DE ASOCIACIÓN O EFECTO). Valor del parámetro que indica ausencia de asociación o efecto entre exposición y enfermedad. Si la medida es la diferencia de riesgos, de tasas o de *odds*, el valor nulo es 0, y si la medida es la razón de riesgos, de tasas o de *odds*, el valor nulo es 1.

