

Cuadernos Gestión Sanitaria

Número 8

**EPIDEMIOLOGÍA
CLÍNICA**



8º Cuaderno de Gestión Sanitaria

EPIDEMIOLOGÍA CLÍNICA

Autor:

Carlos Campillo Artero

Revisión:

Jordi Gol Montserrat
Gema Pi Corrales

FUNDACIÓN



FUNDACIÓN GASPAR CASAL

ISBN: 978-84-09-13190-7

D. L.: M-24599-2019

Maquetación, impresión y encuadernación:

Impresión A2

Crta. M-506 – Km. 25,800 – Vía de Servicio
28320 - Pinto (Madrid)

ÍNDICE

Prólogo	7
Capítulo 1. Introducción a la epidemiología	9
Capítulo 2. Medidas de frecuencia y ajuste de tasas	15
1. Medidas de frecuencia	16
1.1. Definiciones imprescindibles	16
1.2. Algunos ejemplos	18
1.3. La importancia de la variabilidad de las estimaciones	20
2. Ajuste de tasas	23
2.1. Descripción del problema del ajuste	23
2.2. Definición de tasas ajustadas y ejemplos	24
2.3. Métodos de ajuste de tasas	25
2.3.1. Ajuste de tasas por el método directo	25
2.3.2. Ajuste de tasas por el método indirecto	26
2.3.3. Ejemplos de ajustes de tasas y recapitulación	28
2.4. Bibliografía	30
Capítulo 3. Tipos de estudios: ensayos clínicos	31
1. Introducción	32
2. Validez interna y validez externa	32
3. Tipos de estudios: clasificación	32
4. Asignación aleatoria	36
5. Objetivos y diseños	38
6. Enmascaramiento (blinding)	43
7. Flujos de participantes (abandonos, pérdidas)	44
8. Variables (primarias y secundarias)	45
9. Análisis por intención de tratar	46
10. Métodos y mediciones (basal y sucesivas)	47
11. Resultados	47
11.1. Unidad de análisis, medias y varianzas	48
11.2. Análisis de subgrupos	48
11.3. La tiranía de las medias	49
12. Fases de un ensayo	50

13. Consideraciones finales.....	51
14. Material didáctico complementario	52
15. Bibliografía	53
Capítulo 4. Estudios observacionales, medidas de asociación y sesgos	55
1. Estudios observacionales: clasificación	56
2. Estudios de cohortes	56
2.1. Medidas de asociación en los estudios de cohortes y ejemplos	57
3. Estudios de casos y controles	59
3.1. Medidas de asociación en los estudios de casos y controles y ejemplos.....	61
4. Estudios transversales o de prevalencia.....	65
4.1. Medidas de asociación en los estudios transversales y ejemplos	66
4.2. Prevalencia y relación entre medidas de asociación	66
5. Sesgos	67
6. Otras aplicaciones de los estudios observacionales.....	71
7. Bibliografía	72
Capítulo 5. Confusión e interacción.....	73
1. Confusión	74
1.1. Definición y ejemplos	74
1.2. ¿Qué se debe hacer con las variables de confusión?	77
2. Interacción	79
2.1. Definición y ejemplos	79
2.2. ¿Cómo actuar cuando se detecta interacción?	81
3. Confusión e interacción en los modelos multivariantes	82
4. Bibliografía	84
Capítulo 6. Pruebas diagnósticas. Parte I.....	85
1. Introducción.....	86
2. Criterios de validación de las pruebas diagnósticas	86
3. Validez clínica	87
4. Utilidad clínica	93
5. Algunas cautelas	94
6. Material didáctico complementario.....	95
7. Bibliografía	95

Capítulo 6. Pruebas diagnósticas. Parte II.....	97
1. Introducción.....	98
2. Capacidad discriminante y construcción curvas ROC	98
3. Comparación de biomarcadores: capacidad discriminante incremental	102
4. Medidas de asociación, capacidad discriminante y factores de riesgo	105
5. Evaluando la calibración y la capacidad discriminante de modelos multivariantes.....	105
6. Dos últimos ejemplos.....	112
7. Recapitulación.....	114
8. Material didáctico complementario	115
9. Bibliografía	116
Capítulo 7. Asociación, riesgo y causalidad	117
1. Introducción.....	118
2. <i>Association is not causation</i>	118
3. Interpretando asociaciones, riesgo y causalidad	120
4. Explicar y predecir.....	121
5. Causalidad inversa.....	122
6. Estudios que permiten estimar efectos causales en ausencia de aleatorización	122
6.1. Variables instrumentales	123
6.2. Estudio cuasi-experimental pre-post con grupo control (diferencia en diferencias)	126
6.3. Series temporales interrumpidas	128
6.4. Propensity score.....	129
7. Material didáctico complementario	131
8. Bibliografía	132
Referencias del material didáctico complementario	135

Carlos Campillo Artero

Doctor en Medicina y cirugía, Master in Public Health (Johns Hopkins University), Especialista en Medicina preventiva y salud pública. Ex-funcionario y actual consultor de la OMS, responsable de evaluación clínica y de servicios de salud en el Servei de Salut de les Illes Balears, miembro del Comité Técnico Institucional de Seguridad del Paciente, Ministerio de Sanidad, Servicios Sociales e Igualdad, Codirector académico del Máster en Economía de la Salud y Farmacoeconomía e investigador asociado en el Centro de Investigación en Economía y Salud (CRES, BSM, UPF).

PRÓLOGO

Investigar para generar conocimiento que mejore la práctica clínica y el sistema sanitario es una tarea más del buen profesional sanitario. Ello no implica que todos los profesionales tengan que dedicar la misma parte de su tiempo a la investigación, pues hay muchas formas de servir a los pacientes, al sistema y a la sociedad en general. Se puede servir bien dedicando todo el tiempo a la asistencia, todo a la gestión, y casi todo a la investigación (pues nunca hay de quejar la asistencia), según las preferencias y habilidades personales y las condiciones del medio de trabajo. Sin embargo, dedicar algo de tiempo a la investigación es muy satisfactorio con independencia del conocimiento que generemos; es así, porque aunque la mayoría de nosotros seamos solo modestos “obreros de la ciencia”, la investigación nos facilita estar en la frontera del conocimiento en un campo (y ello puede mejorar la calidad asistencial), es un estímulo para seguir aprendiendo muchas cosas, nos permite relacionarnos con algunos de los mejores investigadores, y nos hace sentir parte de un esfuerzo común de mejora del mundo a través de la ciencia. También tiene otros beneficios concretos, como una mayor posibilidad de viajar a reuniones científicas y hacer avanzar una carrera académica.

Para investigar no hay que tener grandes habilidades técnicas. Basta con tener ilusión, acercarse a algún grupo solvente de investigación, y arremangarse al lado de sus miembros para sacar adelante algunas tareas. Hace años se decía que había que “pegarse a un buen investigador”. Hoy la unidad funcional básica en investigación no es un investigador principal sino un grupo, pues la investigación se ha profesionalizado mucho y se ha hecho bastante compleja. Para “sobrevivir” en el ambiente competitivo de la ciencia hay que formar parte de un grupo, deseablemente con varios investigadores principales, que te proporciona recursos, un ambiente estimulante, y muchos conocimientos. Investigar es en buena medida un “oficio” que, al igual que el de pintor o escultor en el Renacimiento y el Barroco, se aprende dentro de un taller (el grupo). Al igual que entonces, ahora hay aprendices que superarán en muchas cosas a sus compañeros y a los investigadores principales (los antiguos maestros) al pasar por varios grupos a lo largo de su periplo profesional. Pero, ni los cursos breves, ni los másters, ni los libros son suficientes para aprender este oficio. En los últimos años evito participar como docente en algunos cursos de metodología de investigación, porque en ellos me siento mal: he conocido a demasiada gente que ha hecho muchos cursos y nunca ha logrado investigar, por falta de un buen grupo o de condiciones locales para hacerlo. Por eso, solo acudo a los cursos que proporcionan a los estudiantes una oportunidad de investigar (y aprender investigando) durante varios meses, dentro de un buen grupo.

Pero para que el aprendizaje sea eficiente, se requiere disponer y haber trabajado con un buen material docente antes de entrar en el grupo y durante la estancia en el mismo. El Cuaderno de Epidemiología Clínica del Dr. Carlos Campillo cumple perfectamente este propósito. Tiene todo lo importante y no le sobra nada, usa buenos ejemplos de estudios

reales bien publicados, y transmite la experiencia acumulada por su autor, que es un excelente investigador y un gran docente. Como tal, alerta de los errores en el diseño de los estudios, en el análisis de los datos y en su interpretación, pero no se ensaña con los errores ni con quienes los cometen. En definitiva, no pretende formar a “cazasesgos”, sino ayudar a los clínicos a convertirse en investigadores sensatos e inteligentes. Este Cuaderno de Epidemiología ayuda a entender el dicho de que el mejor amigo del hombre no es un perro, sino un buen libro. Desde luego, éste lo es (y también se puede leer junto a un perro).

Fernando Rodríguez Artalejo

**Catedrático y Director del Departamento de Medicina Preventiva y
Salud Pública, Universidad Autónoma de Madrid**

Epidemiología clínica

Capítulo 1. Introducción a la epidemiología

Objetivos Generales

1. Aprender los fundamentos y métodos básicos de la epidemiología clínica y sus aplicaciones prácticas en el ámbito de la investigación clínica, de servicios y políticas de salud y de la gestión sanitaria. En este Cuaderno el hincapié recae mucho más en que dichos fundamentos y métodos se entiendan conceptualmente que en el aprendizaje de los métodos cuantitativos.
2. Construir un aparato crítico mínimo que ayude a interpretar y evaluar correctamente los diseños, métodos y resultados básicos de los estudios epidemiológicos, sus aplicaciones y su alcance en los campos de la clínica, la evaluación de servicios y políticas de salud y la gestión sanitaria.

Objetivos específicos de este Capítulo

Adquirir una visión general de la epidemiología como disciplina, de sus objetivos, objetos de estudio, alcance, áreas, nuevas aplicaciones, así como de sus relaciones con otras disciplinas y limitaciones.

Se han publicado numerosas introducciones a la epidemiología excelentes, como, por ejemplo, las que se encuentran en los libros que el correr de los años ha convertido en verdaderos clásicos y que, por este motivo, se han incluido en la bibliografía general de este cuaderno.

Como sería una osadía vana intentar emularlas, esta Introducción se circunscribe a definir la epidemiología, enumerar sus objetivos fundamentales, sus principales áreas de aplicación y, cómo no, sus limitaciones. Se dedicarán, además, unos pasajes a describir de forma sucinta algo importante y relativamente reciente: el enriquecimiento metodológico del que se ha beneficiado esta disciplina en la última década, su procedencia, importancia práctica y el amplio abanico de nuevas aplicaciones y posibilidades que la utilización de nuevos y no tan nuevos diseños de estudios y métodos de análisis estadístico ha abierto y tenemos a nuestra disposición tanto en la esfera de la clínica, como en la de la investigación, evaluación y gestión —macro, meso y micro— de servicios de salud y en la planificación sanitaria.

La epidemiología se ha definido como el estudio de la distribución de las enfermedades en las poblaciones y de los factores que *influyen* en la etiología o *causan* dichas enfermedades y su distribución. Hace bastantes décadas que los padres de la epidemiología moderna se preguntaban por qué unas personas enferman y otras no, por qué unas personas expuestas a los que se consideran factores de riesgo no desarrollan determi-

nada enfermedad y otras no expuestas a ellos la padecen. De lo que no cabe duda es de que la distribución de las enfermedades en la población no es aleatoria, sino que sigue determinados patrones que la epidemiología ayuda a identificar y entender.

En su famoso y ya clásico diccionario de esta disciplina, John M. Last define la epidemiología como el estudio de la distribución de las enfermedades y de los determinantes de la salud en poblaciones concretas y —algo muy importante en este cuaderno— la aplicación de los resultados de este estudio al control de los problemas asociados con la salud (no sólo de las enfermedades).

De esta definición parten la concepción, el diseño y los contenidos de este cuaderno de introducción a la epidemiología. En él se aúnan los objetivos y el contenido temático de esta disciplina con su creciente ámbito de aplicación en muchas y diversas esferas y con distintas ciencias o disciplinas, incluso con algunas alejadas sólo en apariencia de las ciencias de la salud, como la ciencia política, la sociología, la economía o la ingeniería.

Los clásicos señalan cinco objetivos básicos de la epidemiología. Primero, contribuir a conocer la etiología o las causas de las enfermedades y los factores que aumentan el riesgo de padecerlas. Segundo, estimar el alcance, la magnitud o la carga de los problemas asociados con la salud en las poblaciones. Tercero, estudiar la historia natural y estimar el pronóstico de las enfermedades. Cuarto, evaluar, sobre todo, la eficacia, la efectividad y la seguridad de las actuaciones encaminadas a la prevención de la enfermedad y la promoción de la salud y, no menos importante, de las intervenciones en el ámbito de la atención sanitaria y en la prestación de servicios de salud. Quinto, proporcionar datos, información y conocimiento para diseñar, planificar, implantar, desplegar, monitorizar y evaluar políticas, estrategias, planes, programas y todo tipo de servicios y prestaciones de salud, así como para regular y evaluar las tecnologías médicas.

La definida hace pocos años como ciencia regulatoria (*regulatory science*) y la evaluación de tecnologías sanitarias han avanzado mucho merced a la utilización, entre otras herramientas, el de la conceptualización y el arsenal metodológico que usa la epidemiología y otras disciplinas en la toma de decisiones regulatorias respecto, por ejemplo, a la autorización de nuevos medicamentos, dispositivos médicos y pruebas diagnósticas, a las decisiones sobre su priorización, cobertura y reembolso y a la evaluación de su utilización y de su efectividad y seguridad postcomercialización.

Como el *marco* de aplicación de la epidemiología lo conforman la sociedad, las poblaciones, los grupos de población y el individuo, y su *objeto* de estudio es la salud —entendida en el sentido amplio de la definición de la Organización Mundial de Salud—, sus aplicaciones se despliegan en un amplio terreno que, necesariamente, abarca, entre otros, los determinantes sociales de salud, las políticas, no sólo las de la salud, el medio ambiente, la investigación básica y traslacional, la clínica, la prevención, y las políticas y los servicios de salud.

Por estas razones, la epidemiología se puede aplicar y de hecho se aplica (y se debería aplicar más) en el diseño, la planificación, la vigilancia, la gestión macro, meso y micro y la evaluación de servicios de salud, de una gran variedad de políticas que influyen de un modo u otro, directa o indirectamente, en la salud de las poblaciones (de ahí el lema *Salud en todas las políticas*, su significado y sus implicaciones cruciales) y en la investigación de las relaciones entre salud y, por ejemplo, educación, empleo, renta, impuestos, formas de pago a proveedores, seguridad vial, medio ambiente, comercio, así como de los efectos de estas relaciones tanto en grupos como en individuos. Todo ello se produce en un entramado de multicausalidad realmente complejo que apenas se empieza a entender.

En este cuaderno, el arsenal metodológico (diseños, tipos de estudios, análisis estadísticos) que se explica se ha acotado *ex profeso* al de las aplicaciones y la utilidad de esta disciplina en el ámbito de la clínica, la gestión clínica y, también, aunque menos, de la meso y la macrogestión.

Para facilitar la comprensión de los contenidos, teniendo en cuenta que la epidemiología es una asignatura más en masters y diplomas universitarios no especializados en ella, se ha huido de los enfoques eminentemente cuantitativos para centrar la atención en la comprensión conceptual de sus principios, fundamentos y métodos. No es un cuaderno de nivel intermedio ni avanzado, sino una introducción a lo básico de la epidemiología en la esfera de su aplicación en clínica y gestión. Aquí por lo *básico* se entiende aquello que constituye —o debería constituir— una de las herramientas imprescindibles del razonamiento y de la toma de decisiones clínicas y de gestión (adecuadas, efectivas, eficientes) de todos los profesionales de la salud.

Por estos motivos y por la limitación de horas lectivas, centrarse en *lo básico* obliga asimismo a dejar al margen otras aplicaciones de la epidemiología que también pueden interesar al clínico —y a tantos otros—, como la amplia gama de las que pueden apreciarse en salud pública: promoción de la salud, cribado poblacional de enfermedades, estimación del potencial de prevención a escala poblacional, investigación de brotes, evaluación de la efectividad de intervenciones en grupos de población (por ejemplo, efectos del fomento de la actividad física, estilos saludables, ciudades saludables, impuestos y regulación del comercio de determinados alimentos en las enfermedades cardiocerebrovasculares, de la legislación en seguridad vial en la mortalidad...). Quedan también al margen la epidemiología genética, la medioambiental y la aplicada específicamente a diferentes tipos de enfermedades.

En este cuaderno, vale la pena insistir en ello, se han priorizado lo básico y los métodos sobre los contenidos específicos de cada área de aplicación. Sin métodos ni se entiende ni se avanza; con ellos es posible entender y avanzar en cualquier área de interés particular.

Por consiguiente, se ofrece una visión parcial, pero no desactualizada ni sesgada, de las aplicaciones de esta disciplina. Como se apunta al principio de esta introducción, los años recientes y la más candente actualidad están siendo testigos de la incorporación en la epidemiología de un nutrido contingente de diseños de estudios y métodos de análisis cuya

omisión, incluso en un cuaderno introductorio, convertiría sus contenidos en desfasados e incompletos.

Tal como se verá en varios capítulos, el aumento del acceso a inmensas bases de datos y de la capacidad de vincular bases de datos distintas entre sí ha abierto las puertas a información ingente (*record linkage, big data, data mining*). Esto, unido al enriquecimiento metodológico señalado, está ayudando a solventar en parte limitaciones consideradas hace pocos años difícilmente superables.

Nos referimos, por citar solo algunos ejemplos, a relativamente nuevos diseños de ensayos clínicos más eficientes que los tradicionales (diseños *umbrella*, *basket*, adaptativos, *plataformas*, *stepped-wedged*, *pick-a-winner*...). También aludimos a la utilización en medicina de diseños y técnicas de análisis econométricos y de otras áreas de conocimiento, que se utilizan en ellas desde hace décadas, como las variables instrumentales, las series temporales con regresiones segmentadas, los diseños de diferencias-en-diferencias, los grupos control sintéticos, el *propensity score*, y los *regression and classification trees* y los *random forest*, procedentes de la inteligencia artificial y el *machine learning*, con los cuales, a semejanza de los ensayos clínicos, también se pueden investigar relaciones de causalidad y estimar efectos causales, algo muy distinto a probar causalidad.

Virtualmente desconocidos en epidemiología hasta hace pocos años, estos diseños y métodos de análisis se están utilizando cada vez más con datos de la vida real (*real world data*) en estudios observacionales. Así puede comprobarse en una cifra creciente de artículos publicados en revistas médicas de todas las especialidades. Con todos ellos, bien utilizados, tomando las precauciones necesarias por los peligros que puede acarrear su mal uso —pues los estudios observacionales están más expuestos a amenazas a la validez interna que los experimentales—, se está avanzando mucho en varios frentes, como en la construcción y utilización de grupos control adecuados o en la estimación de efectos causales con diseños no experimentales y en condiciones de práctica clínica habitual. Esto puede contribuir, y ya lo está haciendo de forma palpable, a estimar mejor la efectividad y la seguridad, por ejemplo, de nuevos tratamientos en la práctica, con datos recabados por nosotros durante nuestro trabajo cotidiano, y a reducir de este modo la brecha entre eficacia y efectividad, entre validez interna y externa, como se explicará en varios capítulos. Muestras inapelables de ello son los fondos crecientes que se están destinando en muchos países a la investigación de la efectividad comparada (*comparative effectiveness research*) y los resultados que se están obteniendo.

Algunos de estos métodos están siendo muy útiles también para avanzar en el campo de la medicina predictiva, estratificada y personalizada. Primero, a la postre, no interesan tanto los efectos de intervenciones o de medicamentos en grupos —los que se estiman habitualmente con ensayos clínicos en condiciones experimentales, casi ideales, pero más o menos alejados de la práctica clínica—, como dichos efectos en los individuos. La certeza de esta afirmación aumenta cuando entendemos y somos capaces de capturar y medir heterogeneidad de efectos, no sólo entre subgrupos, sino entre individuos, no sólo las medias de los efectos, sino también sus varianzas. Segundo, necesitamos mejorar nuestra capaci-

dad tanto de explicar, como de predecir y discriminar efectos, respuestas en individuos (en qué personas será efectivo y seguro un nuevo medicamento, quién responderá y quién no, en quién será tóxico). La epidemiología ayuda en ambos frentes, como se verá en los Capítulos que siguen. Todo ello se ha hecho insistiendo en la utilidad de emplear métodos y diseños que, aunque se conocen hace muchos años, se utilizan y explican menos de lo necesario.

Como se insistirá en varios apartados de este cuaderno, es sumamente importante tener en cuenta que la clave al utilizar los estudios observacionales para estimar efectos causales es que deben concebirse y diseñarse intentando aproximarlos (remedar) en la medida de lo posible al ensayo clínico aleatorizado y controlado que *sería ideal hacer* pero no puede realizarse por numerosos motivos que conocemos y que nos conducen a los observacionales como única alternativa para estimar efectos causales. La validez de estos estudios observacionales (la de los efectos causales estimados con ellos) estriba en la medida en que con ellos se logre tener grupos de comparación lo más parecidos posible a los que se consiguen con la asignación aleatoria, una tarea difícil, que no siempre se consigue. Esto último subraya la importancia decisiva de comprobar que se cumplen los requisitos específicos de cada estudio observacional en concreto. Aún y así, virtualmente nunca es posible descartar la existencia de confusión residual.

Afortunadamente, el reconocimiento de la importancia y sobre todo de la utilidad en clínica y las limitaciones señaladas de estos métodos y diseños explica que se estén usando cada vez más y mejor, como también puede comprobarse en muchos estudios publicados en los años más recientes. Establecer un diagnóstico con biomarcadores con validez o utilidad clínica bajas o poco conocidas, hacer recomendaciones a partir de factores de riesgo que realmente no lo son o aceptar erróneamente la existencia de una asociación causal entre una intervención clínica u organizativa en los servicios de salud y una mejoría observada en grupos, en pacientes o en indicadores de gestión de unos servicios, además de poder causar daño, es ineficiente o inútil en el mejor de los casos.

Cuando estos hechos se contemplan desde una perspectiva global, social, la de todos, la del paciente que tenemos enfrente, la del siguiente, la de los pacientes que atienden nuestros colegas, la de los que se atenderán en el futuro, así como la que tiene presente que los que hoy no son (somos) pacientes probablemente lo serán (seremos) en algún momento, y cuando, además, tenemos claro que los recursos no son infinitos, es fácil darse cuenta de que los recursos utilizados al establecer un diagnóstico erróneo o sin consecuencias o al administrar un tratamiento inadecuado o innecesario (multiplicados por los n pacientes que nosotros atendemos y los n que atienden los demás) dejan de estar disponibles para hacer otro diagnóstico o administrar otro tratamiento, es decir, tienen un alto coste de oportunidad.

Se ha demostrado que, en el agregado, las cifras de la *inadecuación* son importantes en todo el mundo. La epidemiología ayuda a conocer las variaciones geográficas de dicha inadecuación (por ejemplo, en la prescripción de medicamentos, aparatos, pruebas diagnósticas, cirugía), a estimar sobre e infrautilización (sobrediagnóstico, sobretratamiento), la magnitud del despilfarro (*waste*) o la cuantía y naturaleza de eventos adversos asociados con ella (conocidos y no pocos evitables).

Por consiguiente, la epidemiología brinda información necesaria para diseñar intervenciones dirigidas a mejorar la adecuación de la prescripción, la efectividad de diagnósticos y tratamientos, la seguridad del paciente, la prestación de servicios. Y ayuda, además, a fundamentar mejor las decisiones de los reguladores, gestores y clínicos y, al cabo, los resultados en salud, la eficiencia y la solvencia del sistema de salud.

Aquí se reconoce que en muchos estudios epidemiológicos publicados se cometen errores de diseño, conducción, análisis e interpretación de resultados con su propia distribución de frecuencias: unos errores son muchos más frecuentes que otros. Por algo será. Con el fin de contribuir a reducir su incidencia, se insiste en diversos apartados en el riesgo de cometerlos y se explica la forma de evitarlos.

Muy importante, y por último, en este cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros y convenientes, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I, 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como Material didáctico complementario.

Espero que este cuaderno introductorio, cuyos contenidos se han ido adaptando con el paso de años compartidos con clínicos en el dédalo de los hospitales y en centros de atención primaria y en cursos, sea útil y agradeceré sobremanera todos los comentarios que reciba para seguir mejorándolo.

Bibliografía

Nivel introductorio

- Gordis L. Epidemiology, Second Edition. Philadelphia, PA: WB Saunders Company; 2000.
- Mausner JS, Kramer S. Epidemiology: An introductory text. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. Foundations of epidemiology. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epidemiología clínica. Ciencia básica para la medicina. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. Epidemiology, Beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc.; 2000.

Capítulo 2.

Medidas de frecuencia y ajuste de tasas

Índice

1. Medidas de frecuencia	16
1.1. Definiciones imprescindibles	16
1.2. Algunos ejemplos	18
1.3. La importancia de la variabilidad de las estimaciones.....	20
2. Ajuste de tasas	23
2.1. Descripción del problema del ajuste	23
2.2. Definición de tasas ajustadas y ejemplos	24
2.3. Métodos de ajuste de tasas	25
2.3.1. Ajuste de tasas por el método directo	25
2.3.2. Ajuste de tasas por el método indirecto	26
2.3.3. Ejemplos de ajustes de tasas y recapitulación	28
2.4. Bibliografía.....	30

Objetivos específicos de este Capítulo

1. Conocer e interpretar correctamente las medidas de frecuencia más utilizadas y el significado de la variabilidad e incertidumbre asociadas con sus estimaciones y su importancia en la planificación de intervenciones en clínica y en servicios de salud.
2. Conocer cuándo y por qué deben ajustarse las tasas y aprender los métodos de ajuste directo e indirecto y el significado de las tasas ajustadas.

Nota importante

En este Cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I, 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como **Material didáctico complementario**.

1. Medidas de frecuencia

1.1. Definiciones imprescindibles

Veamos, en primer lugar, seis conceptos básicos, cuyo significado correcto ha desvirtuado el mal uso del lenguaje.

Razón (ratio): es el cociente que resulta de dividir dos cifras, A y B siendo ambas distintas. Por ejemplo:

la razón de sexos = No. de hombres / No. de mujeres en una población y periodo determinados

Odds = probabilidad de que ocurra un evento (p) / probabilidad de que no ocurra el evento (1-p)

Proporción (proportion): es el cociente que resulta de dividir A / A + B. En este caso A está tanto en el numerador como en el denominador, como en los siguientes ejemplos:

$$\text{Proporción de votantes al partido X} = \frac{\text{No. individuos que votan a X}}{\text{Total de la población}}$$

$$\text{Prevalencia de una enfermedad} = \frac{\text{No. individuos con la enfermedad}}{\text{Total de la población}}$$

$$\text{Sensibilidad de una prueba} = \frac{\text{No. verdaderos positivos}}{\text{No. de verdaderos positivos y falsos negativos}}$$

Tasa (ratio): es una proporción que incorpora la dimensión del tiempo. Por ejemplo, la incidencia y la densidad de incidencia, que se definen un poco más abajo.

Prevalencia (prevalence): es el cociente que resulta de dividir los sujetos con determinada característica en una población (**prevalentes**) por el total de dicha población en un periodo acotado. Por tanto, la prevalencia es una proporción, porque el numerador está incluido en el denominador.

$$\text{Prevalencia} = \frac{\text{No. de casos de una enfermedad}}{\text{Total de la población}}$$

Por ejemplo, si en una población de 5000 personas en 2014, 400 tenían diabetes, la prevalencia de diabetes en ella = $400 / 5000 \times 100 = 8,0\%$.

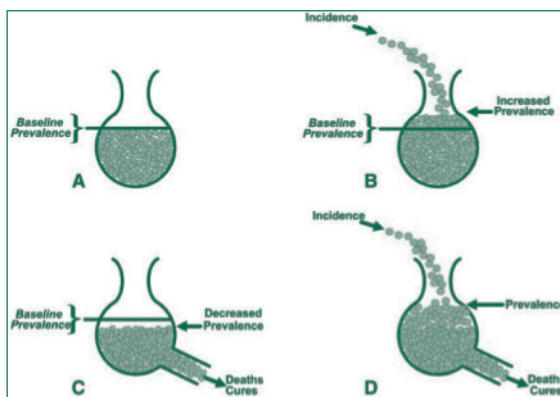
Incidencia (acumulada) (incidence): es el cociente que resulta de dividir el No. de casos **nuevos** diagnosticados (**incidentes**) en un periodo acotado / No. de sujetos en dicha población **en riesgo** de padecer la enfermedad. La incidencia es una tasa, porque su deno-

minador incorpora el período de tiempo dinámico durante el cual van apareciendo casos nuevos.

$$\text{Incidencia acumulada} = \frac{\text{No.de casos nuevos de una enfermedad}}{\text{Total de la población}}$$

Si en aquella misma población en 2014 se diagnosticaron 30 casos nuevos de diabetes, la incidencia de esta enfermedad en ese año = $30 / 5000 (\times 1000) = 6,0 \times 1000$.

Fijémonos en que la prevalencia es un concepto estático, una proporción de casos existentes en un periodo. Por el contrario, la incidencia expresa la incorporación de nuevos casos a una población y por ello es dinámico y expresa, a diferencia de la prevalencia, el riesgo de dicha población de padecer diabetes. Cuanto mayor sea la incidencia, mayor es el riesgo. La siguiente figura ilustra la relación entre prevalencia e incidencia.



Relationship between incidence and prevalence.

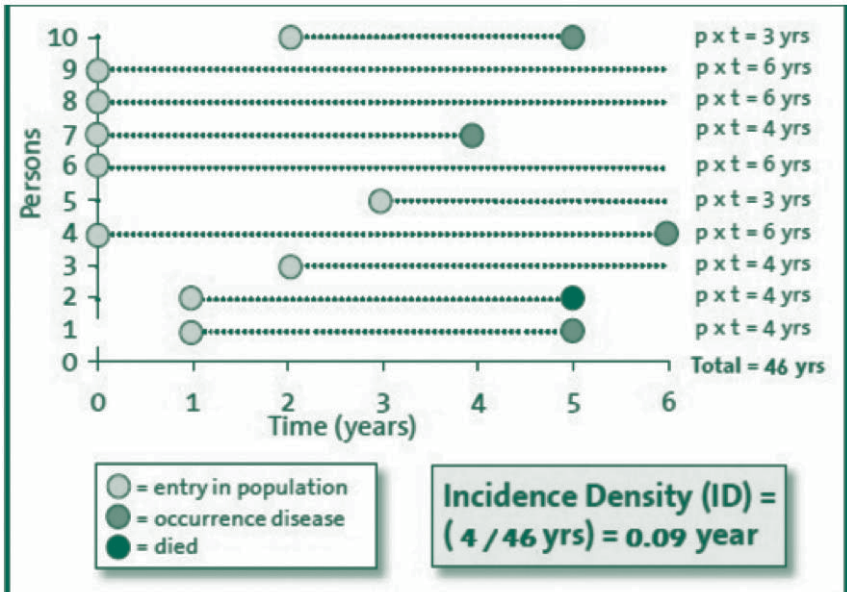
<https://changethru.info/relationship-between-and/relationship-between-incidence-and-prevalence-of-copd.php>

La incidencia, sin embargo, no es un indicador exacto de riesgo, porque hay personas en una población o grupo de estudio que entran y salen de él en diferentes momentos del periodo estudiado y cuyo seguimiento no se ha podido realizar durante todo el periodo acotado al que se circunscribe la incidencia y, además, no todas esas personas están expuestas el mismo tiempo.

La estimación más exacta de la incidencia y, por consiguiente, del riesgo, que no está afectada por estos hechos, es la **densidad de incidencia (*incidence density*)**, que se define así:

$$\text{Densidad de incidencia} = \frac{\text{No.de casos nuevos de una enfermedad}}{\text{Suma de los tiempos de todos los individuos de la población o el grupo de estudio que han estado expuestos}}$$

Veámoslo en la siguiente ilustración:



En el estudio ilustrado en la figura se han seguido 10 personas, en 4 se ha desarrollado la enfermedad (las personas 1, 4, 7 y 10) y 1 ha fallecido (la 2). La suma de los periodos de seguimiento de esas 10 personas ha sido de 46 años y, así, la densidad de incidencia es 4 casos nuevos por 46 años de seguimiento (en total) o 0,09 casos por persona-año de seguimiento.

1.2. Algunos ejemplos

Ejemplifiquemos y analicemos los conceptos descritos con casos prácticos. Tomemos el primer ejemplo de los datos que ofrece el Proyecto Neumonía Zero de la OMS y el Ministerio de Sanidad, dirigido a prevenir las neumonías asociadas con ventilación mecánica en pacientes ingresados en las UCIs (datos de 2008).

Número de neumonías X 100 / total de pacientes: **6,11 %**
 $845 \times 100 / 13.824 = 6,11$ neumonías por cada 100 pacientes

6,11% es una tasa, porque el numerador está contenido en el denominador y porque, además, como se indica en el informe, los casos del numerador son *nuevos* (no son prevalentes, sino incidentes) y están contenidos en el denominador. Es, más concretamente, la *incidencia* de neumonías que desarrollaron los pacientes ingresados en 2008 en las UCIs incluidas en el estudio. Mide el riesgo de desarrollar una neumonía de todos los pacientes ingresados.

Número de neumonías X 100 / pacientes con ventilación mecánica: **13,90 %**
 $845 \times 100 / 6.080 = 13,90$ por 100 pacientes con ventilación mecánica

13,9% también es una tasa por los mismos motivos y esta incidencia es más alta porque del total de pacientes ingresados en las UCIs (denominador de la proporción anterior o 13.824) se han sustraído los que no se encontraban con ventilación mecánica, cuyo riesgo de desarrollar la neumonía es menor. Por consiguiente, como cabe esperar, la incidencia estimada es más alta porque traduce el mayor riesgo de desarrollarla de los intubados.

Número de neumonías X 1000 / total de estancias (tabla mensual de factores)
 $845 \times 1000 / 107.610 = 7,85$ por 1000 días de estancia

7,85 ‰

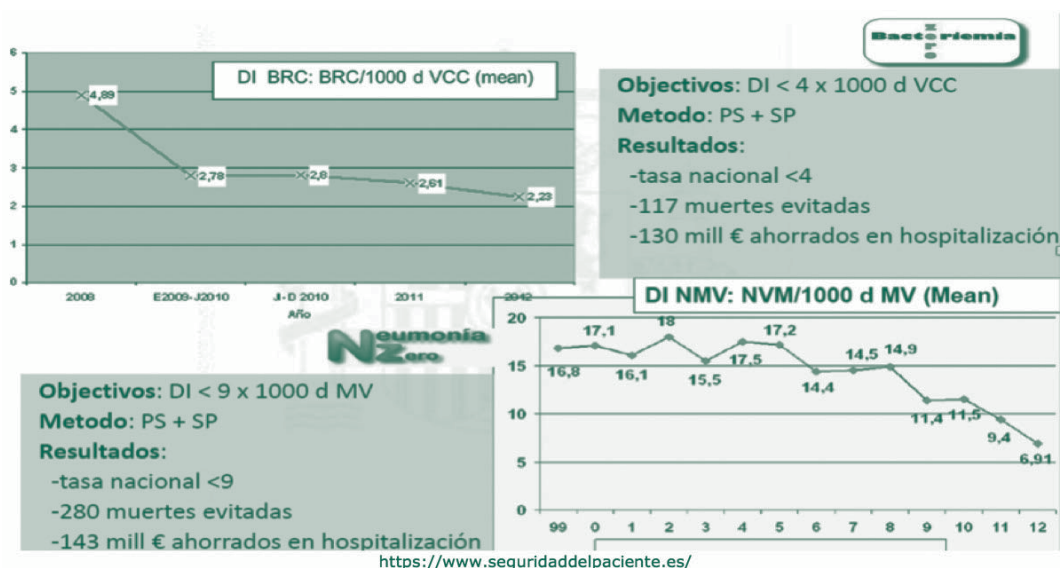
7,85 x 1000 días de estancia también es incidencia, porque los casos del numerador son nuevos, pero en este caso en relación con las estancias. Y, por último,

Número de neumonías X 1000 / total de días de ventilación mecánica
 $845 \times 1000 / 56.527 = 14,95$ por 1000 días de ventilación mecánica

14,95 ‰

14,95 x 1000 días de ventilación mecánica es una tasa, concretamente la *densidad de incidencia*, porque los casos son nuevos (en el numerador) y el denominador incorpora la suma de los tiempos de todos y cada uno de los pacientes que estuvieron expuestos a la ventilación mecánica en el periodo de estudio, es decir, el tiempo total de exposición.

Veamos ahora las siguientes figuras y estimaciones, que corresponden a la evolución de la *densidad de incidencia* (DI) de bacteriemia en pacientes con catéter venoso central (BRC) (bacteriemias *nuevas* diagnosticadas por 1000 días de catéter venoso central colocados) y de neumonía en pacientes con ventilación mecánica (NMV) (neumonías *nuevas* diagnosticadas por 1000 días de ventilación mecánica) en 230 UCIs de hospitales españoles.



El descenso de las DI en ambos procesos traduce una reducción del riesgo de desarrollarlas, un logro importante. Tras esta reducción, que se produce después de la implantación de un conjunto de medidas para prevenirlas, late una hipótesis de causalidad: ¿se debe a las medidas preventivas adoptadas en la atención de estos pacientes o a otras causas?

Valga señalar que los autores del estudio parecen haber establecido esta relación de causalidad e incluso atribuyen a las intervenciones de los proyectos BZ y NZ las muertes evitadas y el gasto evitado como consecuencia de las neumonías y bacteriemias prevenidas.

Cada uno debe valorar por sí mismo en qué medida es aceptable el establecimiento de esta asociación causal. Al finalizar este cuaderno, se podrá responder a esta pregunta sobre causalidad, que es esencial en investigación.

1.3. La importancia de la variabilidad de las estimaciones

Por otro lado, debe tenerse en cuenta que en las curvas anteriores sólo se incluyen las estimaciones puntuales y globales de la densidad de incidencia (DI) “media” de todas las UCIs sin las medidas de incertidumbre asociadas con ellas (los errores estándar a partir de los cuales se construyen los IC95% de las DI) ni de dispersión (cómo varían las densidades entre UCIs).

Como se verá en otros capítulos, las medias son importantes, pero, de cara a diseñar y planificar intervenciones y tomar decisiones, más importante es aún conocer la variabilidad asociada en torno a estas estimaciones medias puntuales, la heterogeneidad de densidades de incidencia entre las UCIS, cómo y cuánto varían entre UCIs. Cabe esperar que haya UCIs con DI más altas y otras con DI más bajas que la media estimada. Sin medidas de variación (varianza, desviación estándar, error estándar) no es posible conocer la magnitud real del problema en cada UCI de las estudiadas ni diseñar intervenciones específicas adaptadas y adecuadas para cada una de ellas. La heterogeneidad, la varianza es más importante que la media.

De hecho, incluso en las UCIs grandes, con unidades de traumatología, neurocríticos, coronarias, general, etc., centrarse exclusivamente en la DI de bacteriemia media en un año esconde algo frecuente en estas UCIs, que se deduce al analizar la siguiente tabla (extraída del Estudio ENVIN: <http://hws.vhebron.net/envin-helics/>) y resta información clave para diseñar actuaciones efectivas.

PATOLOGIA DE BASE	N	%	Estancia Media	APACHE II	Exitus (%)	Densidad de incidencia (*)
Coronaria	35	0,25	20,60	20,55	45,71	2,71
Médica	313	2,26	28,96	20,66	32,27	7,90
Quirúrgica	174	1,26	28,59	18,61	22,41	7,68
Traumatológica	139	1,01	27,09	15,91	10,79	15,62

Como se muestra en la tabla, la magnitud del problema (Apache I, mortalidad, densidad de incidencia) varía notablemente en las diferentes secciones de esta UCI (compárese, por ejemplo, la DI de la unidad coronaria con la de traumatología), lo que no muestra la DI media de 8,48 bacteriemias por 1000 días de catéter venoso central. Obviamente, las acciones de mejora que pueden diseñarse conociendo o desconociendo las DI de cada UCI pueden ser distintas y en algunos casos inefectivas e ineficientes. Como se explicará en los próximos capítulos, intervenir a partir del conocimiento exclusivo de medias omitiendo la heterogeneidad cuando ésta existe tiene sus peligros y se conoce como el problema de la *tiranía de las medias*.

Fijémonos ahora en las medidas de frecuencia de un último ejemplo, correspondiente al Estudio nacional de eventos adversos ligados a la hospitalización (ENEAS) de 2006:

- IC95% de la incidencia de eventos adversos = 8,6-10,1
- Densidad de incidencia (DI) = 1,4/100 días de estancia
- 16% fueron graves
- 38% se atribuyen a errores de medicación
- 25% a infecciones nosocomiales
- 25% a procedimientos
- 43% eran evitables

Si lanzamos sobre esta información una mirada desde el prisma de la epidemiología, podemos hacer varias consideraciones. La incidencia y la densidad de incidencia nos ofrecen estimaciones del riesgo de padecer eventos adversos durante la hospitalización. En el primer caso, al contrario que con la DI, se omite la estimación puntual, pero el IC95% nos aporta más información, porque incluye el recorrido de valores de la incidencia congruentes con los datos observados con una confianza del 95%.

Para conocer la magnitud global del problema también deberíamos conocer las variaciones de estas estimaciones en los diferentes hospitales, como comentamos en el ejemplo de las UCIs.

Las causas a las cuales se atribuyen los eventos adversos hemos de revisarlas a la luz del concepto de causalidad. En muchos casos, no cabe duda del papel causal que desempeña un determinado factor, pero en otros es sumamente difícil, por no decir imposible, establecer con certeza la causalidad, al igual que puede ocurrir al afirmar que un evento adverso concreto era evitable. Estos temas se tratarán en detalle en los restantes capítulos.

Por ejemplo, en una fractura de fémur en un paciente hospitalizado a veces puede ser difícil recabar toda la información necesaria para confirmar si fue patológica o se debió a que resbaló fortuitamente o a que no se habían colocado suelo antideslizante y agarraderas (y en este último caso que fuera evitable). Este caso concreto ha sido uno de los motivos de controversia sobre la lista de eventos adversos evitables cuyo coste de atención algunos sistemas de salud decidieron no reembolsar a los hospitales (por ejemplo, el programa *Do not pay* en los Estados Unidos).

Los porcentajes de las causas de los eventos adversos (38%, 25% y 25%) nos permiten jerarquizar su frecuencia, pero en ausencia de sus numeradores y denominadores no podemos conocer la magnitud real del problema, es decir, si son muchos o pocos casos: 16% de eventos adversos graves o mortales de 100 eventos adversos totales son 16, lo cual no es un problema de la misma magnitud que 16% de eventos adversos graves de un total de 1000 detectados en el estudio (160 eventos adversos graves o mortales). De ahí que sea esencial conocer los numeradores y los denominadores, es decir, las cifras absolutas antes de extraer ninguna conclusión.

A este respecto, valga recordar un error que se comete con frecuencia al trabajar con porcentajes. El descenso del 20% al 15% de cualquier medida no es un descenso del 5%; es un descenso de 5 puntos porcentuales, pero del 25% (5 de 20 es el 25% de 20).

2. Ajuste de tasas

2.1. Descripción del problema del ajuste

Cuando una población se puede dividir en estratos es posible comparar las proporciones (que por mal uso del lenguaje se llaman tasas) específicas de cada estrato, como, por ejemplo, las tasas de mortalidad estratificadas por sexo, edad o por cualquier otra variable. Una **tasa bruta** (que, de hecho, es una proporción) es una *media ponderada* de las tasas específicas de cada estrato que componen la población. En ella, los pesos de ponderación de cada estrato corresponden al tamaño de la población de cada estrato. (Insistimos, en este apartado se utiliza el término de tasa al uso, aunque a sabiendas de que no es correcto en sentido estricto según las definiciones de la sección anterior. La tasa de mortalidad es, en rigor, la proporción de mortalidad; el tiempo no está contemplado en ella.)

La diferencia entre las tasas brutas globales de dos poblaciones (por ejemplo, de mortalidad) puede ser la resultante de las diferencias entre 1) las tasas específicas de cada estrato y 2) la composición (el tamaño) de sus poblaciones en cada estrato. Estas diferencias pueden confundir o distorsionar las **comparaciones de las tasas brutas** de dos poblaciones y, por este motivo, dichas comparaciones pueden no ser adecuadas.

Veamos en el siguiente ejemplo la población 1 estratificada en tres grupos de edad con el número de personas y el de fallecimientos de cada grupo de edad. La tasa bruta de mortalidad es 30 muertes / 300 habitantes = 10 muertes por 100 habitantes:

Population 1			
Stratum i (Age group)	Age (years)	Total population (N _i)	Deaths (X _i)
1	0–4	100	5
2	5–14	90	10
3	15–19	110	15
Total		300	30

Fijémonos en la población 2:

Population 2			
Stratum i (Age group)	Age (years)	Total population (N _i)	Deaths (X _i)
1	0–4	165	10
2	5–14	75	10
3	15–19	60	10
Total		300	30

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/FundEpi/coursePage/lectureNotes/>

Su tasa bruta de mortalidad es también 30 muertes / 300 habitantes, es decir, 10 por 1000 habitantes, como la de la población 1, pero tanto el tamaño de cada grupo de edad como el número de muertes difieren de los de la población 1.

En las dos tablas que aparecen a continuación se aprecia que el tamaño total de las poblaciones es el mismo (300) y que la población 2 tiene una población más joven y tasas de mortalidad específicas de cada estrato más altas que las de la población 1:

Population 1					Population 2				
Stratum i	Age	Total	%	Death rate per 100	Stratum i	Age	Total	%	Death rate per 100
1	0-4	100	33	5	1	0-4	165	55	6
2	5-14	90	30	11	2	5-14	75	25	13
3	15-19	110	37	14	3	15-19	60	20	17

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/FundEpi/coursePage/lectureNotes/>

Las tasas de mortalidad específicas de cada estrato (aquí son x 100 habitantes), que expresan el riesgo de muerte de cada estrato, ya no son iguales como lo eran las tasas brutas (globales): estas tasas estratificadas son más altas en la población 2.

La diferencia en la **estructura de edad** de las dos poblaciones *confunde* la comparación de las tasas brutas globales: el riesgo global de muerte las dos poblaciones es el mismo, pero el riesgo de muerte de cada estrato es diferente, mayor en los tres estratos de la población 2 y, por esta razón, las tasas brutas no son comparables. Para serlo, deberemos *eliminar el efecto* distorsionador de la edad y esto se hace *ajustando por edad* la tasa bruta de mortalidad, como se explica un poco más abajo.

2.2. Definición de tasas ajustadas y ejemplos

Con las **tasas ajustadas** (también se denominan **estandarizadas**) por una variable (por ej., edad) o más variables simultáneamente (edad, sexo, gravedad, renta) se pueden comparar tasas globales de poblaciones entre las cuales existen diferencias en determinadas características o variables (edad, sexo, nivel socioeconómico, morbilidad, gravedad...) que pueden alterar la comparación del efecto estudiado, como ocurre en el caso anterior con la estructura diferente de la edad sobre la mortalidad.

Otro ejemplo: si en un servicio clínico la fracción de casos de edad muy avanzada, graves y con comorbilidad es más alta que la de otro servicio, sus tasas brutas de mortalidad no se pueden comparar con las de otro; para poderlas comparar habrá que eliminar el posible efecto distorsionador de la edad, la gravedad y la comorbilidad en la mortalidad ajustando

sus tasas de mortalidad por estas tres variables (en este caso el ajuste se realizaría por todas estas variables simultáneamente con métodos multivariantes). Las tasas brutas son menores en un servicio porque el riesgo de muerte de sus pacientes (más jóvenes, con menos comorbilidad y menos graves) es menor, no porque sea necesariamente mejor su atención o porque la atención del servicio que tiene las tasas brutas más altas sea peor. Para poder compararlos, hay que ajustar las tasas de mortalidad por esas variables y, al hacerlo, las tasas ajustadas permiten responder a la pregunta: ¿qué mortalidad tendrían los dos servicios (tasas ajustadas de mortalidad) si sus pacientes tuviesen edades, comorbilidades y gravedades similares?

Las tasas ajustadas no tienen significado per se, son artefactuales (resultado de un tratamiento estadístico) y solo permiten comparar tasas de dos poblaciones o grupos porque eliminan la diferente distribución de variables confusoras (edad, gravedad, complejidad de la casuística...), que afectan el acontecimiento de interés: mortalidad, complicaciones, readmisiones, etc. Sólo tienen valor para fines de comparación entre ellas.

Siempre que se comparen tasas ajustadas, antes hay que comparar las tasas brutas y las específicas por estrato y observar la composición de las dos poblaciones respecto de la variable supuestamente confusora (por ej., la edad) respecto a 1) los tamaños de la población de cada estrato y 2) las tasas específicas de cada estrato (como las de mortalidad en el ejemplo anterior).

2.3. Métodos de ajuste de tasas

Los ajustes de tasas pueden realizarse mediante varios métodos:

- Comparando tasas por estratos en el análisis estratificado
- Con los métodos directo e indirecto de ajuste de tasas, como se explica a continuación
- Por medio de tablas de vida (no se explicarán en este cuaderno)
- Con análisis multivariante (que permite ajustar simultáneamente por más de una variable)

2.3.1. Ajuste de tasas por el método directo

Para entender cómo se ajusta por el método directo, utilicemos el siguiente ejemplo. En él se presentan las tasas de mortalidad de cada grupo de edad de México y de los EEUU. Hemos averiguado que los tamaños de los estratos de cada población son diferentes y que la tasa bruta (global) de mortalidad de México es 4,7 por 1.000 habitantes y la de los EEUU, 8,7 por 1.000 habitantes. Observamos que las tasas de mortalidad por grupos de edad de México son más altas que las de los EEUU. En otras palabras, la estructura de edad (la población de los EEUU es más vieja que la de México) y la “experiencia” de mortalidad (las tasas de mortalidad de cada estrato) de ambas poblaciones son diferentes: el riesgo de muerte de todos los grupos de edad de México es más elevado que el de los correspon-

dientes grupos de edad de los EEUU, pero la tasa bruta de mortalidad de México es menor. Una aparente paradoja. Es posible, por consiguiente, que la edad esté *distorsionando* la comparación de tasas y, por esta razón, hemos de ajustar las dos tasas brutas por edad para eliminar este efecto de la edad en la comparación. Además, en la tabla se incluye una **población estándar** (por ejemplo, la del mundo en el mismo periodo).

	Población estándar (1)	Tasa de mortalidad específica por edad por 100.000 habitantes, 1995-1997		Número esperado de muertes	
		México (2)	Estados Unidos (3)	México (4) = (1)x(2)/100.000	Estados Unidos (5) = (1)x(3)/100.000
<1	2.400	1.693,2	737,8	41	18
1-4	9.600	112,5	38,5	11	4
5-14	19.000	36,2	21,7	7	4
15-24	17.000	102,9	90,3	17	15
25-44	26.000	209,6	176,4	55	46
45-64	19.000	841,1	702,3	160	133
65+	7.000	4.967,4	5.062,6	348	354
	100.000			639	574

Boletín Epidemiológico / OPS, Vol. 23, No. 3 (2002)

Para ajustarlas, multiplicamos las tasas específicas de cada estrato de México y de los EEUU por separado ((2) y (3), que conocemos) por la población estándar (de referencia) de cada estrato (1), para obtener el número de muertes *esperadas* en cada grupo de edad de México y de los EEUU que cabe esperar si sus respectivas poblaciones tuvieran la misma estructura de edad que la del mundo. Las **tasas ajustadas** de México y de los EEUU se calculan dividiendo las muertes totales *esperadas* en cada una de ellas (639 y 574) por el tamaño de la población estándar (100.000).

Así, la tasa de mortalidad ajustada por edad de México = (639 / 100.000) x 1.000 = 6,39 por 1.000 habitantes y la de los EEUU = (574 / 100.000) x 1.000 = 5,74 por 1.000 habitantes.

Con el método directo aplicamos las tasas de mortalidad de cada grupo de edad de nuestras poblaciones a la población estándar. Las tasas ajustadas por edad son las que tendrían nuestras dos poblaciones si su estructura de edad fuese la de la estándar, es decir la misma. Si tuviesen la misma estructura de edad, la tasa de mortalidad de los EEUU sería más baja (5,74 por 1.000) que la de México (6,39 por 1.000). La diferencia entre las tasas brutas y las ajustadas por edad demuestra que la edad estaba *distorsionando* la comparación entre las tasas brutas.

2.3.2. Ajuste de tasas por el método indirecto

En este caso procedemos de otro modo. Imaginemos que sabemos que la tasa bruta de mortalidad de Colombia en 1999 fue 4,4 por 100.000 habitantes, con variaciones entre 1,8 por 100.000 en el departamento de Vichada y 6,9 en el de Quindio. (Podemos hacer este ejercicio, por ejemplo, con la tasa de mortalidad de hospitales y las de sus servicios.) Queremos saber si la tasa de mortalidad global de Colombia se puede comparar con la de sus departamentos, por ejemplo, con la de Vichada.

Sin embargo, en este caso (como ocurre con frecuencia) no tenemos las tasas de mortalidad por grupos de edad del departamento de Vichada, sino solamente las muertes en cada grupo de edad (3), el tamaño de cada grupo de edad (2) y las tasas de mortalidad del país por grupos de edad (1).

	Tasas de mortalidad específica por grupos de edad, por 100.000 habitantes, Colombia, 1999 ¹ (1)	Población del departamento de Vichada ² (2)	Muertes observadas en Vichada, 1999 ¹	Muertes esperadas en Vichada (3)
0-4	339	11.392	61	39
5-14	34	21.930	5	7
15-44	219	38.244	27	84
45-64	752	7.083	22	53
65 y más	4.573	1.839	27	84
		80.488	142	267

Boletín Epidemiológico / OPS, Vol. 23, No. 3 (2002)

Para ajustar por el método indirecto, multiplicamos las tasas específicas de cada estrato de la **población estándar**, de Colombia (columna 1), por el tamaño de cada estrato de la población de Vichada (columna 2), para obtener el número de muertes *esperadas* en cada estrato de ella (es decir, cuántas muertes esperaríamos observar en Vichada si tuviesen la experiencia de mortalidad de Colombia). Seguidamente dividimos el número de muertes *observadas* en nuestra población (que conocemos) por el de las *esperadas* y obtenemos así la llamada **razón de mortalidad estandarizada** (RME o *standardized mortality ratio*, SMR).

La RME de Vichada = 142 / 267 = 0,53. Esto significa que el riesgo de muerte en Vichada sería 47% (100 – 53) menor que el que cabe esperar en Colombia según la experiencia de mortalidad de este país ajustando por edad.

Si la RME = 1, el riesgo (de muerte en este ejemplo) es el mismo en nuestra población y en la estándar.

Si la RME <1, el riesgo de muerte es menor en nuestra población que en la estándar.

Si la RME > 1, el riesgo de muerte es mayor en nuestra población que en la estándar.

Si multiplicamos la RME por la tasa bruta de la población estándar, obtenemos la tasa ajustada por el método indirecto.

En otras palabras, **en el método de ajuste directo multiplicamos las tasas específicas de nuestra población por los tamaños de cada estrato de la población estándar. En el indirecto, a diferencia del directo, las tasas específicas por estrato de la población estándar las multiplicamos por el tamaño de cada estrato de nuestra población.**

Tengamos muy presente que, **cuando se utiliza la misma población estándar para calcular las RME de dos o más poblaciones, la RME calculada de cualquiera de ellas no**

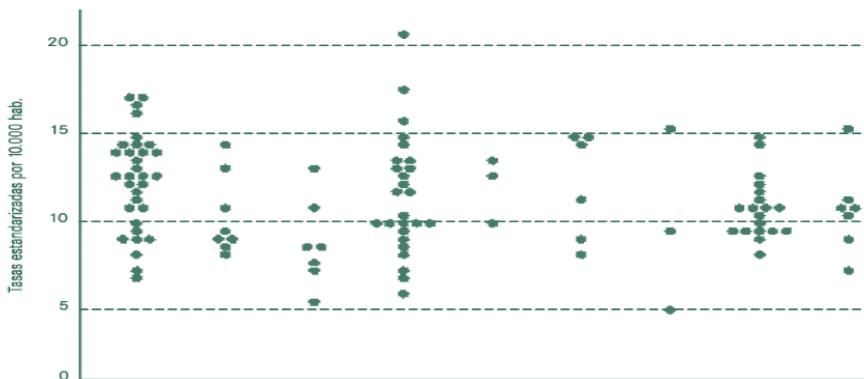
se puede comparar con la de las otras a menos que las dos (o más) poblaciones que se comparen tengan una distribución muy similar de la edad. La RME calculada de cada población sólo puede compararse con la tasa de mortalidad de la población estándar.

2.3.3. Ejemplos de ajustes de tasas y recapitulación

Cualquier tasa ajustada por cualquier variable por este método se interpreta del mismo modo. Veamos algunos ejemplos.

Unas tasas de complicaciones o de mortalidad o de readmisiones ajustadas por riesgo mediante el método indirecto (ICAR, IMAR, IRAR) de nuestro hospital = 1,9 significan que las muertes, complicaciones y readmisiones **observadas** en nuestro hospital son 90% **más elevadas que las esperadas**, que tenemos un 90% más de muertes, complicaciones y readmisiones que el estándar (el conjunto de miles de altas combinadas en una sola población de todos los hospitales que componen la población estándar). En el ICAR, IMAR e IRAR las tasas se han **ajustado simultáneamente** (mediante modelos multivariantes) **por riesgo**, entendiendo el riesgo como un conjunto de numerosas variables tales como, por ejemplo, en el caso de la IMAR, edad, sexo, riesgo de muerte del diagnóstico principal, riesgo de muerte del diagnóstico secundario con máximo riesgo, riesgo de muerte del procedimiento con máximo riesgo, circunstancia de admisión (urgente/no urgente), tipo de GRD (quirúrgico/no quirúrgico), nivel del hospital, ámbito rural o urbano, etc. Es decir, las tres tasas ajustadas por numerosas variables simultáneamente usando métodos multivariantes (ICAR, IMAR e IRAR) son las que tendría nuestro hospital si se comportara en gestión y tuviese la misma población que el estándar respecto a esas variables de ajuste.

Revisemos un último ejemplo, tomado del atlas de variaciones en cirugía ortopédica y traumatología en el Sistema Nacional de Salud, donde se presentan las tasas de intervenciones de fractura de cadera (ajustadas por edad y sexo, por 10.000 habitantes) en las áreas de salud de 9 Comunidades Autónomas (cada punto es un área de salud), y veamos la interpretación que hacen los autores de las tasas ajustadas (eje de las Y).



http://www.atlasvpm.org/documents/10157/22342/atlasnumero1_%281.20MB%29.pdf

Las tasas crudas (brutas) utilizadas representan el número de altas generadas tras la intervención quirúrgica en 2002 (numerador de la tasa) por la población de un área (denominador). Dado que la edad y el sexo son dos variables determinantes de la morbilidad (de especial importancia en las patologías asociadas a la osteoporosis o la artrosis), las diferencias en la distribución de estas dos variables de las poblaciones entre las diferentes áreas justificarían la variabilidad en sus respectivas tasas de intervenciones (por ejemplo, cabe esperar que se intervenga más en áreas con poblaciones más envejecidas). Para controlar este efecto, se calcularon las tasas específicas (estratificadas) de intervención quirúrgica que presentaban los grupos de edad y sexo de las diversas áreas. Para obtener una sola medición que facilitase la comparación entre áreas obviando sus diferencias demográficas, se calcularon (por el método directo) las tasas estandarizadas de intervención por edad y sexo, empleando como población de referencia (estándar) la de la población española registrada en el censo del 2001 estándar. Estas tasas estandarizadas (ajustadas) pueden no coincidir con las tasas brutas, ya que representan las tasas de intervención quirúrgica que tendrían las diversas áreas si todas tuvieran una población con la distribución de edad y sexo de la población española.

Y tomemos nota de los comentarios que hacen los autores en este caso práctico sobre el ajuste indirecto que también han realizado y de sus similitudes con lo afirmado anteriormente.

Cálculo del número de casos esperados (estandarización por el método indirecto) y de la razón de incidencia estandarizada (equivalente a la RME)

La precisión estadística de los cálculos mencionados varía con el tamaño de la población y el número de intervenciones realizadas. De este modo, el error en la estimación de dichas tasas puede llegar a ser importante para procedimientos con escasos efectivos (pocas intervenciones) o en áreas con poca población, situaciones en las que es preferible analizar las razones de incidencia estandarizada (RIE), que requieren estimar los casos esperados de intervención en cada área (método indirecto de ajuste).

El número de casos esperados (de intervenciones esperadas) en cada área se obtuvo calculando, para el conjunto de la población estudiada, las tasas específicas de intervención por grupo de edad y sexo, multiplicando estas tasas (riesgo de intervención) por el tamaño de la población de cada área, y sumando los resultados. Se obtiene así el número de intervenciones que cabría esperar en las diversas áreas si cada uno de sus subgrupos de edad-sexo se comportara siguiendo el patrón de la población general (estándar en este caso). En otras palabras, el número de intervenciones esperadas representa el número de intervenciones que se realizarían si los habitantes de las respectivas áreas tuvieran una probabilidad de ser sometidos a tales intervenciones similar a la de la población general de la misma edad y sexo (método de ajuste indirecto).

A diferencia del método directo, y dado que ahora se aplican unas tasas específicas constantes (las de la población estándar) sobre las pirámides de edad de cada área, este método no permite la comparación (de las RME) entre áreas, pues no pueden obviarse las diferencias en estructura de edad y sexo entre ellas, pero sí permite la comparación de (la tasa de) cada

área con la (tasa de la) población estándar. La estandarización indirecta es un método usualmente empleado cuando se desconocen las tasas específicas locales (cuando solo se conoce el número absoluto de sucesos) o cuando estas tasas de cada estrato se basan en muestras pequeñas y, por tanto, imprecisas, como ocurre en algunos de los procedimientos revisados en el Atlas como las intervenciones de espalda y hombro.

2.4. Bibliografía

General

Nivel introductorio

- Gordis L. Epidemiology, Second Edition. Philadelphia, PA: WB Saunders Company; 2000.
- Mausner JS, Kramer S. Epidemiology: An introductory text. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. Foundations of epidemiology. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epidemiología clínica. Ciencia básica para la medicina. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. Epidemiology, Beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc.; 2000.

Específica de este Capítulo

En las referencias del nivel introductorio y en la del intermedio hay capítulos que tratan estos temas con mayor detalle y ejemplos.

Capítulo 3.

Tipos de estudios: ensayos clínicos

Índice

1. Introducción	32
2. Validez interna y validez externa.....	32
3. Tipos de estudios	32
4. Asignación aleatoria	36
5. Objetivos y diseños.....	38
6. Enmascaramiento	43
7. Flujos de participantes (abandonos, pérdidas).....	44
8. Variables (primarias y secundarias).....	45
9. Análisis por intención de tratar	46
10. Métodos y mediciones (basal y sucesivas)	47
11. Resultados.....	47
11.1. Unidad de análisis, medias y varianzas.....	48
11.2. Análisis de subgrupos	48
11.3. La tiranía de las medias	49
12. Fases de un ensayo	50
13. Consideraciones finales	51
14. Material didáctico complementario.....	52
15. Bibliografía	53

Objetivos específicos de este Capítulo

1. Conocer los estudios experimentales (ensayos clínicos), las principales características de los objetivos, tipos de diseños, realización, análisis e interpretación de los resultados de los ensayos clínicos.
2. Conocer las fases de desarrollo de los ensayos clínicos.
3. Familiarizarse con los errores y las limitaciones más frecuentes de los ensayos clínicos publicados.

Nota importante

En este Cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I, 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos

conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como **Material didáctico complementario**.

1. Introducción

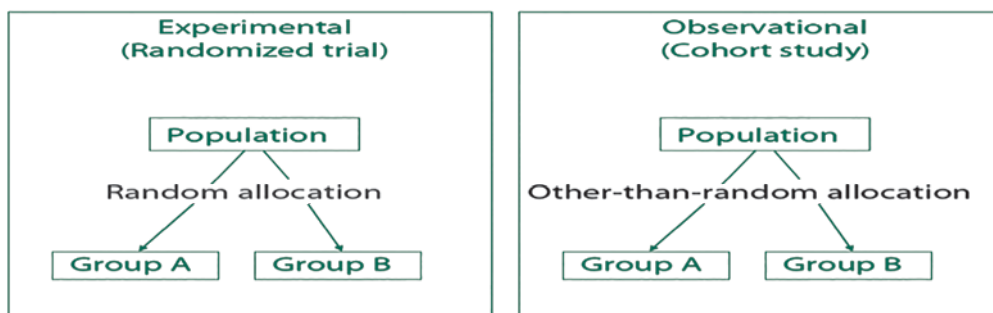
Si bien los ensayos clínicos, como tantos otros tipos de estudios, abarcan un tema muy amplio y con novedades muy frecuentes en las esferas metodológica, regulatoria, clínica, comercial, de la innovación, de políticas de salud y gestión de servicios, en este cuaderno sólo se abordan, a título introductorio, los aspectos metodológicos relacionados con ellos: fundamentos, diseño, realización, análisis e interpretación de resultados.

2. Validez interna y validez externa

Estos son dos conceptos fundamentales que hacen referencia a cualquier tipo de estudio, sea experimental, observacional o descriptivo. La **validez interna** se refiere a la medida en que el diseño, la realización, el análisis y la interpretación y presentación de resultados son adecuados desde el punto de vista del rigor científico y de los principios metodológicos vigentes. La **validez externa**, por su parte, hace referencia a la medida en que los resultados de un estudio individual pueden extrapolarse a personas, muestras o a cualquier otro tipo de participantes de lugares y circunstancias distintas a las de dicho estudio, por ejemplo, a los pacientes que cada uno atiende en su servicio o centro de salud.

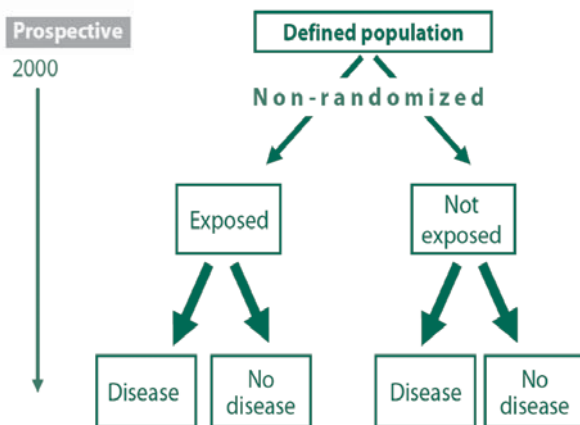
3. Tipos de estudios: clasificación

Para distinguir los tipos de estudios más usados en epidemiología hemos de utilizar tres criterios. El primero es la forma en que se asignan los participantes a los grupos del estudio. El segundo, si el investigador interviene (introduce modificaciones) o no en ellos. Si se realiza mediante asignación aleatoria y se interviene, nos encontramos con estudios **experimentales** (el ensayo clínico). Si la asignación no es aleatoria y no se interviene, con **observacionales**. Si la asignación no es aleatoria y el investigador interviene, los estudios se llaman **cuasi-experimentales**.



<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

El tercero es la direccionalidad: si el estudio parte de los expuestos y no expuestos a la supuesta causa, los sigue en el tiempo hasta que en ambos grupos se observa una respuesta o efecto, se trata de estudios **prospectivos** (*prospective*). Si primero se seleccionan casos (individuos con la enfermedad o la característica de interés) y controles (individuos sin ellas) y entonces se retrocede en el tiempo para averiguar en ambos grupos quiénes estuvieron expuestos y no expuestos, trabajamos con estudios **retrospectivos** (*retrospective*). No debe confundirse esta direccionalidad (de exposición a enfermedad y viceversa) con la dirección que se sigue en el calendario al realizar el estudio, como se verá más adelante. Cuando la temporalidad (exposición y resultado) no desempeña ningún papel, es decir, la exposición y los resultados se observan al mismo tiempo, los estudios se denominan **transversales** (*cross sectional*).



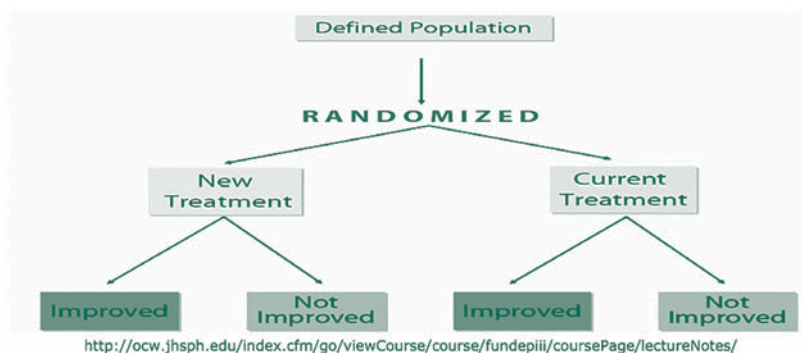
<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundeplii/coursePage/lectureNotes/>

A diferencia de los estudios observacionales, en los cuales no se interviene, en un estudio experimental los investigadores intervienen activamente en la historia natural de un proceso (administrando un nuevo tratamiento para una enfermedad, desplegando una intervención de salud, social o urbana, implantando una política para evitar accidentes de tráfico, imponiendo un copago, etc.) y alterando algunas variables bajo condiciones prediseñadas y controladas (administrando un medicamento) con objeto de hacer inferencias sobre la presencia de efectos, relaciones o asociaciones (de naturaleza causal) entre la exposición a estas intervenciones y sus efectos (respuestas) observados.

A diferencia, también, de los estudios observacionales, en los cuales la construcción de los grupos de estudio no la decide el azar, en los experimentales los grupos (experimental y control) se forman mediante asignación aleatoria; por este motivo se llaman aleatorizados (*randomized*). En estos estudios sólo se emplea la asignación aleatoria para decidir qué participantes forman parte de cada uno de los grupos del estudio.

Por consiguiente, un **ensayo clínico aleatorizado** es un estudio experimental porque en él se interviene en **grupos de intervención y control** formados sobre la base de la asignación

aleatoria de los participantes a ellos. Se llaman **estudios controlados** porque disponen de uno o más grupos control. Los integrantes del grupo experimental (personas expuestas a algún factor o que padecen alguna enfermedad o presentan alguna característica clínica relevante y reciben el nuevo tratamiento experimental, o individuos, grupos de población o poblaciones en los cuales se interviene) **se exponen** a una intervención clínica en el medio clínico-asistencial controlado, reciben el medicamento nuevo (o un dispositivo médico o se someten a una prueba diagnóstica o a una intervención quirúrgica) estudiado, y los de los grupos control, se exponen a otra distinta, reciben otro medicamento (u otra tecnología médica) o placebo. Tras un periodo de seguimiento prefijado, se comparan los resultados observados (los efectos, los resultados, en inglés, los *outcomes* o *end points*) con el fin de hacer inferencias y estimar los efectos de los medicamentos respecto a su eficacia y seguridad en cada grupo (el esquema que sigue muestra el diseño general más frecuente –el diseño paralelo– de estos estudios).



Es muy importante tener claro que **la finalidad del grupo control es poder estimar qué le habría pasado a los participantes del grupo experimental (muy parecidos a los del control) si no se les hubiera administrado el tratamiento o no hubieran recibido la intervención experimental, la que fuere. También se le conoce como *contrafactual*.** Como se verá en el siguiente apartado, con la asignación aleatoria se intentan conformar grupos (brazos, *groups*, *arms*) de pacientes virtualmente idénticos respecto a la distribución de las variables *observadas* y *no observadas* potencialmente asociadas con (o pronósticas o predictivas de) los resultados o el efecto de la intervención (por ej., edad, sexo, comorbilidad, gravedad). **Si la distribución de estas variables que podrían estar asociadas con o ser predictivas de los resultados es muy similar en los dos grupos, las diferencias que se observen en los resultados entre ambos grupos no se deben a ellas, sino a lo único que diferencia a los grupos, el medicamento administrado, la intervención experimental llevada a cabo, lo que realmente nos interesa.**

Si con la asignación aleatoria se consigue una distribución similar de dichas variables en los grupos, la **inclusión del grupo control** es otra *conditio sine qua non* para poder estimar los efectos o resultados en el grupo experimental del medicamento administrado a sus pacientes y no a otros factores (conocidos o desconocidos), es decir, para poder **estimar efectos causales (no para confirmar causalidad, como se verá en el Capítulo**

7) de la exposición (el medicamento, la intervención): resultados observados (la eficacia y seguridad que se intentan estimar). Si lo único distinto entre ambos grupos es que unos reciben el medicamento experimental y otros el placebo u otro medicamento, las diferencias observadas entre ambos grupos cabe atribuirlos al medicamento experimental (a la intervención realizada), no a otra cosa porque en lo demás son muy similares (suponiendo que no ha variado durante el tiempo que haya durado el ensayo).

Deduzcamos lo que ocurriría si, por ejemplo, en el grupo de comparación (control) el porcentaje de pacientes con comorbilidades, de más edad, más graves y en etapas más avanzadas de la enfermedad fuese sustancialmente mayor que en el grupo de intervención y supongamos que todas ellas están asociadas con las variables de respuesta (de eficacia) escogidas en el ensayo: probablemente responderían *menos* al medicamento o a la intervención que recibieran de lo que habrían respondido si la distribución de esas variables hubiese sido muy similar a la del grupo de intervención y, como consecuencia, al responder menos, la respuesta observada (y el efecto estimado) en el grupo intervenido es mayor que en el grupo control y queda sobreestimada al compararlas.

Valga de entrada avanzar que es sumamente importante recordar siempre que la ausencia de asignación aleatoria puede considerarse la principal limitación de los estudios observacionales, porque sin ella no es fácil disponer de grupos de estudio (grupo experimental o de intervención y grupo control) en los cuales la distribución de las *variables observadas (observables)* y *no observadas (unobservables)*, porque no se hayan medido o porque se desconozcan) asociadas con los resultados sea similar y permita, entonces, hacer comparaciones entre los grupos para estimar efectos causales de las exposiciones a intervenciones en las cuales no intervengan sesgos ni factores de confusión, dos conceptos esenciales que se explicarán en los Capítulos 4, 5 y 7.

Hay numerosas guías y listas de factores y criterios para evaluar la validez interna y la calidad de un ensayo clínico, pero, en síntesis, y sin excepciones, los **elementos que se consideran esenciales para evaluarla**, que se van a describir a continuación, son los siguientes:

- Objetivos
- Diseños
- Asignación aleatoria
- Enmascaramiento
- Flujos (abandonos y pérdidas)
- Variables primarias y secundarias
- Análisis por intención de tratar
- Métodos y mediciones (basal y sucesivas)
- Resultados (análisis subgrupos)
- Conclusiones (validez interna y externa)

4. Asignación aleatoria

A pesar de que se demuestra que con la asignación aleatoria se consiguen grupos con una distribución de las variables muy similar, no existe una garantía absoluta de que con alguna asignación aleatoria no se alcance dicha finalidad. A veces, por azar, los grupos formados pueden presentar desequilibrios importantes en la distribución de alguna o de algunas variables asociadas con (o pronósticas de) los resultados o desenlaces.

Para averiguar si se han producido estos posibles desequilibrios con las variables observadas (conocidas que hemos medido), debe revisarse la que habitualmente se presenta como “**Tabla 1**” en la sección de Resultados de los artículos publicados, que muestra cómo ha quedado la distribución de las variables tras la asignación aleatoria. ¿Se “portó bien” el azar? Veamos un ejemplo en la tabla I de un ensayo fase III para el tratamiento del glioblastoma, que muestra que las distribuciones de las variables en ambos grupos son muy parecidas y no hay desequilibrios. Esta tabla se encuentra en el artículo: [Stupp R, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol. 2009;10:459–66.](#)

Por el contrario, en la Tabla 1 del artículo [Harrison CN, et al. Ruxolitinib versus Best Available Therapy for ET intolerant or resistant to Hydroxycarbamide in a Randomized trial. Blood. 2017;26;30\(17\):1889-97](#) aparece el resultado de la asignación aleatoria de los pacientes con trombocitemia esencial a los dos grupos de tratamiento del ensayo clínico, el de los pacientes del grupo control, que recibieron el mejor tratamiento disponible (*best available treatment*: los utilizados con mayor frecuencia en este ensayo fueron hidroxycarbamida, anagrelida e interferón), y el de los del grupo experimental, tratados con ruxolitinib. Los autores señalan que la distribución de las variables en ambos grupos es muy parecida con dos salvedades: en el grupo que recibió ruxolitinib la duración de la enfermedad fue mayor y la hemoglobina, más baja.

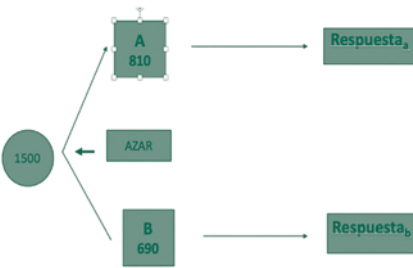
Si comprobamos que estas dos variables (u otras variables cualesquiera en cualquier ensayo) se distribuyen de forma desigual en los grupos de estudio y se trata de una diferencia importante, además de las *medidas brutas* de efecto o de asociación (de resultado) que estemos utilizando en el ensayo, se han de calcular *las medidas de efecto o de asociación (de resultado) ajustadas*. Las medidas de efecto ajustadas resultantes se interpretan como las que se hubieran obtenido si la distribución de esa o esas variables hubiese sido muy similar en ambos grupos de tratamiento.

Recordémoslo siempre: en ausencia de aleatorización no es posible saber con absoluta certeza si alguna variable *no observada* realmente asociada con los resultados se distribuye de forma muy parecida o no en ambos grupos con las consecuencias que ello tiene y hemos comentado en párrafos anteriores. Lo que sí sabemos es que la aleatorización maximiza la posibilidad de formar grupos muy parecidos, que esto hay que comprobarlo, y que, en su ausencia, como ocurre con los estudios observacionales, lo que debe hacerse es minimizar esas consecuencias, algo que veremos cómo puede lograrse en los Capítulos 4, 5 y 7, cuando ya tendremos claro el papel que desempeñan los sesgos y los factores de confusión. El edificio de la epidemiología hemos de construirlo paso a paso.

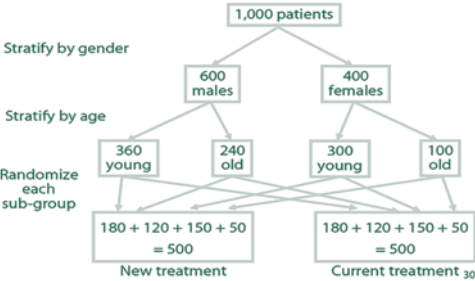
Otro ejemplo: si la presencia de una mutación desconocida (no observada) reduce la respuesta de una persona con cierta enfermedad a un medicamento que se está ensayando y en el grupo control la fracción de participantes con ella, por azar, es más alta que en el grupo experimental, responderán más casos del grupo experimental y menos controles y se sobreestimarà la eficacia del medicamento. Esta eficacia “aumentada” que se observaría no se debe al medicamento experimental, sino a que al comparar las respuestas en los dos grupos, los del control (con más mutaciones) responden menos.

Los **métodos de asignación aleatoria** más frecuentes son la simple, la estratificada y por bloques (se esquematizan en las siguientes figuras), si bien hay otros más complejos, que obligan a hacer análisis estadísticos más avanzados y trascienden los objetivos de este cuaderno.

Asignación aleatoria simple

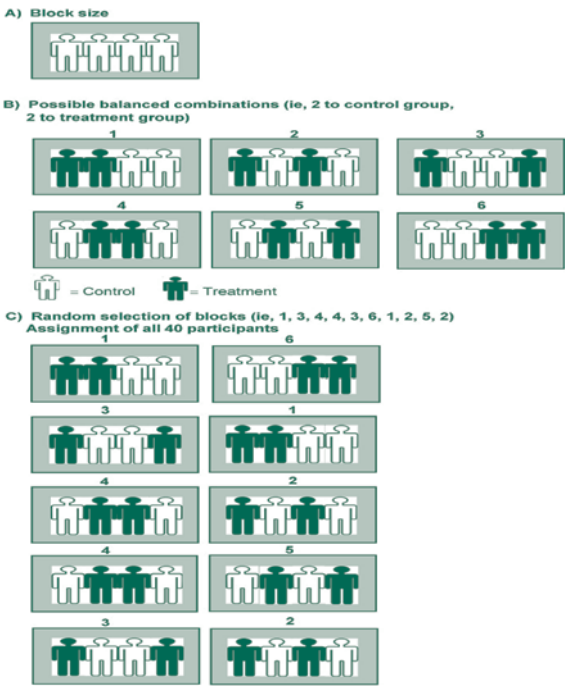


Asignación aleatoria estratificada



<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

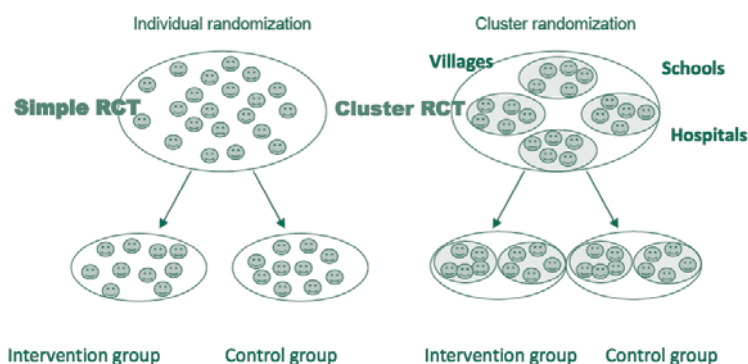
Asignación aleatoria por bloques



Cada método de aleatorización tiene sus ventajas e inconvenientes y requiere métodos de análisis específicos.

<https://pdfs.semanticscholar.org/76a8/94a7f31f671125792589418911fc63574599.pdf>

En la asignación aleatoria por bloques no se asignan individuos uno a uno a los grupos del ensayo, sino que la unidad de asignación (aleatorización) son grupos de individuos. Una variante utilizada de este método son los ensayos aleatorios por conglomerados (**cluster randomized trials**) en los cuales la unidad de asignación son conglomerados (como vecindarios, hospitales, escuelas, pueblos...), como se aprecia en las siguientes figuras. Estos diseños, al igual que sus análisis y la interpretación de sus resultados, revisten mayor complejidad que la de los paralelos por numerosos motivos, como, solo por citar un ejemplo, la correlación entre variables (correlación intraclase) que se observa con frecuencia entre los individuos que pertenecen a un mismo conglomerado porque comparten características por el mero hecho de dicha pertenencia común (Fuente: www.cochrane-handbook.org).



<https://pdfs.semanticscholar.org/76a8/94a7f31f671125792589418911fc63574599.pdf>

5. Objetivos y diseños

En términos prácticos, el objetivo principal de los ensayos clínicos, estimar la eficacia y la seguridad de un medicamento, acostumbra plantearse con una de las siguientes dos finalidades:

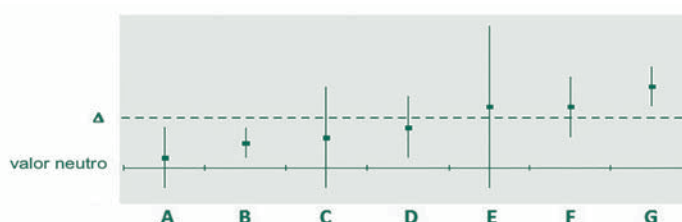
- Estimar si la eficacia y la seguridad de un medicamento son **mayores** que las de otro o frente a placebo (**ensayos de superioridad**, el diseño más frecuente) o
- Estimar si son **iguales o solamente un poco inferiores** (con un umbral de no inferioridad prefijado con criterios clínicos) a las de otro medicamento activo (**ensayos de equivalencia o de no inferioridad**). La frecuencia de estos ensayos está aumentando en los últimos años.

Ambos, si su diseño es paralelo, con uno o más grupos control, se realizan tal como se ha ilustrado anteriormente. Sin embargo, en los de equivalencia o no inferioridad se acuerda (como se muestra en el esquema que sigue) un margen de no inferioridad (conocido como delta) por debajo del cual el medicamento experimental se considera inferior y dentro del cual se considera equivalente o con menor eficacia, aunque aceptable (no inferior). Suelen realizarse cuando el fin es obtener un medicamento de “similar”, aunque sea algo menor, eficacia que otro existente, pero, por ejemplo, con menos efectos secundarios o parecido

en eficacia y seguridad aunque administrado una vez al día en lugar de varias o por una vía de administración diferente, más cómoda y que mejore la adherencia. También se usan para demostrar equivalencia de genéricos y, a veces, para sacar al mercado medicamentos equivalentes a otros disponibles (conocidos como me too), de hecho sólo con fines comerciales, como, por ejemplo, aumentar la competencia, aunque de hecho muchas veces no la aumenten.

En los siguientes dos esquemas se representan las estimaciones puntuales de eficacia (puntos) con sus IC95% (líneas asociadas con los puntos) para ilustrar situaciones de superioridad, equivalencia o no inferioridad, e inferioridad.

No inferioridad:



A juzgar por los IC95% de las estimaciones representadas en este esquema, sólo G debe considerarse inferior, porque tanto la estimación puntual del efecto (cuadrado negro) como su IC95% están por debajo del umbral delta de no inferioridad. B es claramente equivalente y no inferior: su IC95% está dentro del umbral entre el valor neutro y delta. No puede descartarse que C, D, E y F puedan ser inferiores, dado que hay valores de sus IC95% por debajo de delta. Asimismo, A, C y E también pueden ser superiores, puesto que sus IC95% tienen valores por encima del valor neutro.

La eficacia y la seguridad de un medicamento estimadas frente a placebo se denominan eficacia y seguridad **absolutas**. Cuando se estiman en relación con un comparador activo, se llaman **relativas** o **incrementales**. Debe quedar claro cuál y cómo es el **comparador** o, si el ensayo tiene más de un brazo, los comparadores. Si el comparador es placebo, la eficacia y seguridad estimadas para el medicamento experimental (el nuevo) serán mayores que si se compara con el medicamento disponible de mayor efectividad y más seguro. El comparador puede ser placebo o un medicamento activo, y en este último caso, por ejemplo, la molécula más parecida o el medicamento con mayor eficacia o efectividad disponible o uno de eficacia intermedia o el más utilizado con la misma indicación. En cualquier caso, lo imprescindible es saber qué tipo de comparador se utiliza por la importancia que tiene esta información para averiguar si la eficacia del medicamento nuevo (del grupo experimental) se puede estar sobre o subestimando debido al tipo de comparador utilizado.

Un ejemplo de baja eficacia relativa es la mostrada en la figura 1 del artículo [Fojo T, Mailankody S, Andrew Lo A. Unintended Consequences of Expensive Cancer Therapeutics—The Pursuit of Marginal Indications and a Me-Too Mentality That Stifles Innovation and](#)

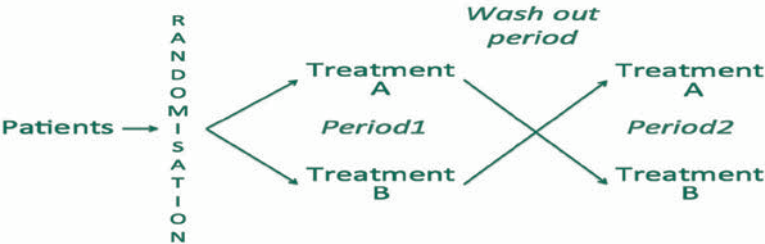
CreativityThe John Conley Lecture. JAMA Otolaryngology–Head Neck Surg. 2014;140:1225-36. En ella las ganancias (incrementos) en eficacia relativa se estiman por medio de 1) la supervivencia libre de progresión (o PFS; la mediana es poco más de dos meses y se debería conocer el límite inferior de su IC95%) y 2) la supervivencia global (OS, la mediana es 2,1 meses) de los medicamentos mencionados.

Existen otros tipos de diseños de ensayos clínicos, más complejos que el paralelo, que aquí sólo se describen sucintamente. Cuando uno se encuentra con uno de los siguientes diseños, lo más recomendable es pedir asesoramiento a un colega que los conozca bien, porque tanto el diseño, como la realización, los análisis y la interpretación de resultados se tornan más complicados.

Veamos algunos a vuelapluma. En el **diseño factorial** se estima la eficacia y seguridad de dos medicamentos por separado (A y B) y conjuntamente (A+B) frente a placebo y frente a cada uno de ellos por separado, como se muestra en el siguiente esquema y se describe en el artículo [Freidlin B, et al. Two-by-Two Factorial Cancer Treatment Trials: Is Sufficient Attention Being Paid to Possible Interactions? JNCI. 2017;109. doi: 10.1093/jnci/djx146.](#)

		Medicamento A	
		No	Sí
Medicamento B	No	Placebo o tratamiento estándar	A
	Sí	B	A +B

En los diseños cruzados (*cross-over*), en una primera fase se realiza un ensayo paralelo. Tras un periodo de lavado (*wash out period*), durante el cual se espera a que se eliminen los medicamentos del organismo, se produce el entrecruzamiento (*cross-over*) y los pacientes de un grupo reciben el medicamento administrado al otro grupo en el periodo anterior. El periodo de lavado, determinado con información farmacocinética, ha de ser suficiente para evitar que la respuesta observada al final no sea el resultado del efecto esperado con el segundo medicamento más el del primero que aún no se ha eliminado (efecto de arrastre, *carry over effect*), un efecto que, si existiese, podría ser el resultado de antagonismo entre los dos medicamentos a las dosis remanentes en el momento de la medición o sinérgico (y en este caso, aditivo o multiplicativo).

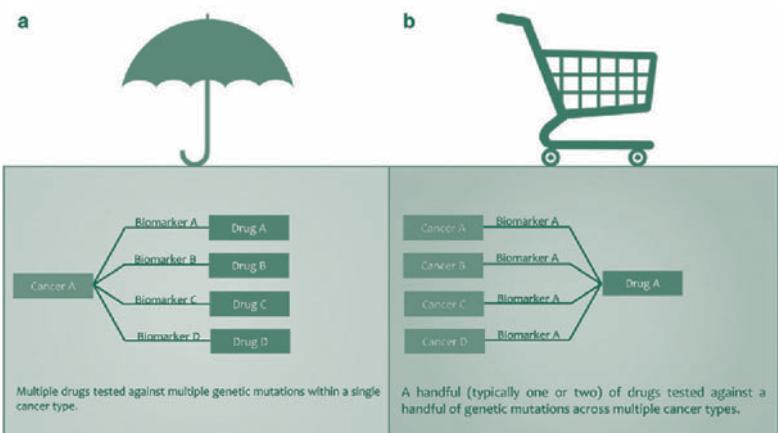


Hay ensayos cuyo diseño, realización y análisis son mucho más complicados, como los ensayos en cuña por pasos (**stepped-wedge designs**) representados a continuación. Se realizan por fases: en la primera todos los grupos son controles, se avanza en el tiempo cambiando a los grupos paulatinamente al tratamiento experimental hasta que en la última fase todos los grupos reciben el tratamiento experimental.

	T0	T1	T2	T3	T4
Cluster 1	Control	Intervention	Intervention	Intervention	Intervention
Cluster 2	Control	Control	Intervention	Intervention	Intervention
Cluster 3	Control	Control	Control	Intervention	Intervention
Cluster 4	Control	Control	Control	Control	Intervention

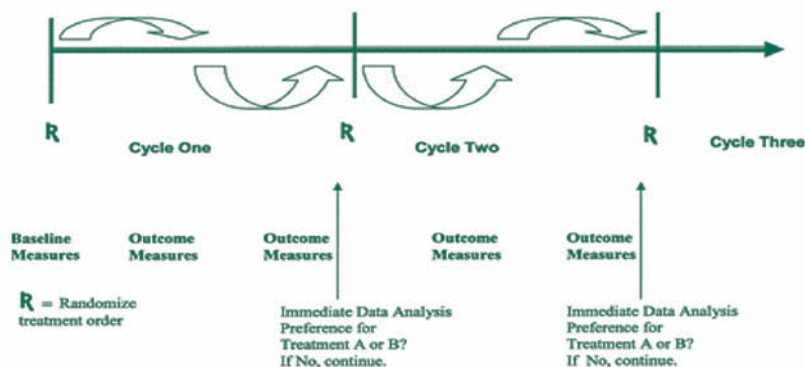
Fig. 1 Design of the stepped-wedge trial

O los **paraguas** (*umbrella*), **cesta** (*basket*) y **plataformas** (*platforms*), esquematizados en las siguientes figuras. Los paraguas se emplean para estudiar varios tratamientos distintos con diferentes dianas terapéuticas en una sola enfermedad (o tipo de neoplasia). Los cesta se utilizan para estudiar un solo medicamento dirigido a una misma diana terapéutica que se expresa en varias enfermedades o neoplasias o subgrupos distintos de pacientes con una misma enfermedad. Y las plataformas sirven para estudiar varias dianas terapéuticas en una sola enfermedad de forma continuada y permiten retirar e incorporar medicamentos sobre la base de los resultados que se van obteniendo y algoritmos de decisión. En estos nuevos diseños se utilizan **biomarcadores** de respuesta y toxicidad para seleccionar pacientes (ver Capítulo 6) cada vez más frecuencia en el contexto de la **medicina de precisión, estratificada y personalizada**. El artículo de Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. NEJM.2017;377:62-70 ofrece una explicación y varios esquemas excelentes y más detallados.



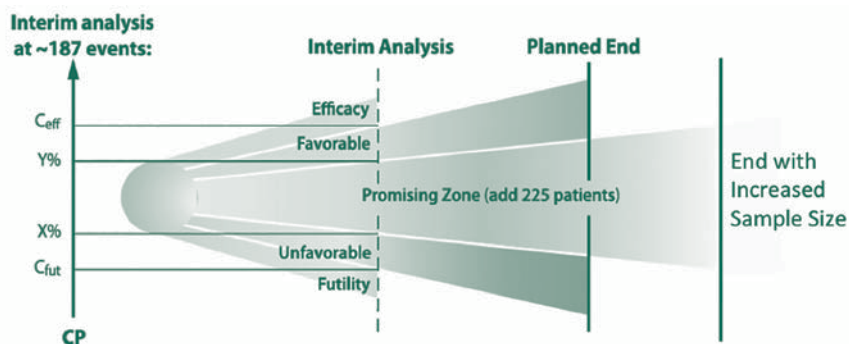
<https://www.semanticscholar.org/paper/Evolving-landscape-of-tumor-molecular-profiling-for-Syn-Yon-g/61e75666bc9dd6e15036b6be6f3bb86bab16bef9>

Otro ejemplo son los **diseños n-de-1** (*N-of-1*), con un solo paciente, al cual se entrecruzan (*cross-over*) por fases dos medicamentos. Estos ensayos se realizan, por ejemplo, con medicamentos indicados para enfermedades raras (huérfanos). En ellos, el mismo paciente ejerce de control de sí mismo y los efectos que se estiman son a nivel individual. (Véase el apartado La tiranía de las medias más adelante en este mismo Capítulo.)



El último tipo de diseño que se comenta en este Capítulo, porque su empleo va en aumento paulatinamente (hay muchas más variantes, aunque se utilizan mucho menos que las descritas), es el de los **diseños adaptativos** (*adaptive trial*), probablemente el más versátil y eficiente, aunque mucho más complejo desde el punto de vista del diseño, la conducción y el análisis, de todos los descritos.

En suma, en estos ensayos se utiliza un diseño *adaptativo* en el sentido de que dicho diseño permite modificar o *adaptar* parámetros clave del ensayo a medida que se observan resultados en dichos análisis intermedios. Estos parámetros son, por ejemplo, el tamaño muestral, las dosis de medicamentos, los pacientes que se van incluyendo y excluyendo en el transcurso del ensayo, aquellos en los cuales los análisis intermedios (*interim analyses*) muestran claramente que el tratamiento no es eficaz o que, por el contrario, es claramente eficaz y seguro, la transferencia de pacientes a grupos del ensayo diferentes de los que fueron inicialmente asignados aleatoriamente porque se dispone de información que permite esperar mejor respuesta en ellos, etc. Lo ilustra el siguiente esquema general:



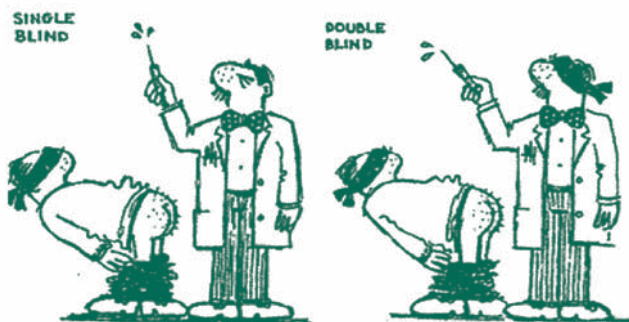
CP = Conditional power
The probability of success (statistical significance) at the end of the trial given current data trend

<http://www.pharmagellan.com/blog/making-clinical-development-more-investable-with-financially-adaptive-trials>

Analizando este esquema de abajo arriba, se aprecia que, tras un primer análisis intermedio (línea vertical discontinua), hay pacientes que no responden (*futility*, en ellos el tratamiento no es eficaz y no causa daño) y abandonan el ensayo, otros (*unfavorable*) en los cuales no es eficaz o causa daño o ambas cosas, y también abandonan el ensayo, otros con resultados promisorios (*promising zone*, parece ser eficaz y seguro) y a él se añaden más pacientes preseleccionados —posiblemente con biomarcadores y criterios clínicos— muy similares a los estudiados hasta ese momento, otros con resultados favorables aunque tampoco definitivos (*favorable*), que también continúan en el ensayo, y un último grupo en cuyos pacientes se estima que el medicamento es claramente eficaz y seguro y abandonan el ensayo como si éste hubiera finalizado, algo que no ocurre, porque al tratarse de un subgrupo el tamaño muestral todavía es insuficiente respecto al planificado a *priori* al diseñar el ensayo.

También lo muestran unos esquemas sumamente ilustrativos de un ensayo adaptativo publicado en la NEJM (Park JW, et al. [Adaptive Randomization of Neratinib in Early Breast Cancer N Engl J Med 2016;375:11-22.](#)), en el cual se compara la quimioterapia neoadyuvante estándar combinada con el inhibidor de la tirosinquinasa neratinib para el tratamiento del cáncer de mama, que se acompaña de una descripción de cada esquema (préstese especial atención al uso de biomarcadores en la selección de pacientes, un tema que se describirá en detalle en el Capítulo 6):

6. Enmascaramiento (blinding)



Para aumentar la capacidad de evitar sesgos (incluso de tipo cognitivo), en los ensayos se utiliza el enmascaramiento hasta que, al finalizar el ensayo, se revelan los códigos de ocultamiento:

- Simple ciego: el paciente no sabe qué tratamiento recibe (intenta evitar que el mero conocimiento del tratamiento que uno recibe o de sentirse observado pueda modificar consciente o inconscientemente las respuestas).
- Doble ciego: ni el paciente lo sabe ni los profesionales que lo atienden conocen qué tratamiento están administrando (a qué grupo del ensayo) y así se evita que se preste una atención diferente a un grupo que a otro (*work-up bias*).
- Triple ciego: lo desconocen el paciente, los profesionales y el personal que analiza los datos. Así se intenta evitar que los datos se analicen de forma distinta, preferencial, en un grupo o en otro.

En ciertos ensayos no es posible enmascarar a estos tres niveles (por ej., el doble ciego en cirugía) y en otros el método de enmascaramiento utilizado puede fallar. En esto radica la importancia de conocer qué métodos de enmascaramiento se utilizaron y, al finalizar el ensayo, si fueron o no efectivos.

7. Flujos de participantes (abandonos, pérdidas)

La evaluación rigurosa de un ensayo clínico exige saber, de los pacientes que lo iniciaron, cuántos acabaron el ensayo, cuántos abandonaron (*withdrawals*) y cuántos se perdieron en el seguimiento (*lost to follow up*).

Además, es crucial saber por qué: los resultados no deben interpretarse del mismo modo si una fracción de pacientes lo abandonaron por discrepancias con la forma en que los atendían, porque padecieron efectos secundarios importantes que obligaron a suspender el tratamiento, porque, por ejemplo, tuvieron que mudarse de país y el hospital del lugar de destino no participaba en el ensayo, o porque algunos tuvieron un accidente de automóvil y tuvieron que abandonar el ensayo (en este caso, incluso, el accidente pudo ser fortuito o la consecuencia de un efecto secundario del tratamiento recibido).

Por todo ello, actualmente es cada vez más frecuente leer notificaciones de ensayos clínicos que incluyen un diagrama de flujo de pacientes en los grupos de estudio donde se indican las pérdidas y abandonos y se especifican sus motivos, como lo ilustra el del ensayo con ruxolitinib y el mejor tratamiento disponible citado anteriormente ([Harrison CN, et al. Ruxolitinib versus Best Available Therapy for ET intolerant or resistant to Hydroxycarbamide in a Randomized trial. Blood. 2017;26;30:1889-97](#)).

La tabla 2, extraída de ese mismo ensayo, es un buen ejemplo de notificación adecuada de los cambios que alteran los flujos de pacientes de cada grupo del ensayo, pues en el estudio se cuantifican y describen el número de veces que se han producido (y los motivos), por ejemplo, cambios en el tratamiento recibido por algunos pacientes de alguno de los grupos (*switches*), la paralización de la participación de pacientes por cambios sustanciales en la enfermedad durante el ensayo (*transformation*, en este ensayo a mielofibrosis, leucemia mieloide aguda o síndrome mieloproliferativo), falta de respuesta, falta de eficacia, toxicidad, muerte, abandonos.

Dicho sea de paso, las **recomendaciones CONSORT**, consensuadas a escala intencional y adoptadas por un número creciente de revistas médicas, son un conjunto de normas justificadas que debe cumplir la notificación de los ensayos clínicos en ellas con el objetivo de incluir en los artículos la mínima información necesaria para que un lector avezado pueda evaluar su validez interna y externa. El fin ulterior que se persigue es reducir las deficiencias de notificación de información esencial de los ensayos publicados, una de las deficiencias principales de estos artículos, así como de otros en que se notifican otro tipo de estudios (de pruebas diagnósticas, estudios observacionales, revisiones sistemáticas, evaluaciones económicas, etc.) en los cuales dicha deficiencia también es de

las más frecuentes. Todas las guías para revisar y publicar cada uno de estos tipos de estudios se encuentran en equator-network.org. Lo más habitual en los artículos publicados es *pecar por defecto*, no por exceso, de información.

En relación con la notificación de información esencial sobre seguridad (al margen ahora de la relativa a la eficacia), existe una ampliación de estas recomendaciones CONSORT dedicada específicamente a esta información, al igual que otra ampliación publicada más recientemente sobre ensayos de equivalencia y no inferioridad, los ensayos paralelos con más de un grupo control. (En la Bibliografía al final de este Capítulo se incluyen las referencias específicas.)

8. Variables (primarias y secundarias)

La elección de las variables más adecuadas para estimar la eficacia y la seguridad de los medicamentos en los ensayos clínicos es una cuestión de naturaleza eminentemente clínica. Es el conocimiento clínico fundamentado el factor que debería decidir en buena medida qué variables miden mejor lo que se tiene que medir, qué variable o variables son las más idóneas, para medir, por ejemplo, la eficacia de un medicamento. La variabilidad de opiniones respecto a cuáles son las variables más adecuadas explica parte de la heterogeneidad de las variables utilizadas en los ensayos realizados con los mismos fines y medicamentos.

A continuación, se hacen varias apreciaciones importantes sobre las variables que pueden encontrarse en los ensayos y sus consecuencias al interpretar los resultados. No debe perderse de vista que al evaluar ensayos siempre debe juzgarse si la estimación de la eficacia y de la seguridad es correcta y razonable o si se incurre en ellos en sobre o infraestimación de una o de ambas.

Siempre se ha de intentar aclarar qué variables son las primarias y cuáles las secundarias, así como si su elección, como se ha indicado, ha sido apropiada. En algunos ensayos se ha comprobado que las variables que se notifican como primarias y secundarias no eran las consideradas así en el diseño y el protocolo del ensayo y que ello responde al interés (no deseable) de sobreestimar eficacia y seguridad: a veces se notifican como primarias aquellas con las cuales se obtienen mejores resultados de eficacia y seguridad, aunque en el diseño a priori fuesen secundarias.

En otros casos, y se trata de una práctica en aumento, las variables escogidas no son finales, del resultado clínico último esperado, sino intermedias, *sustitutivas* (de proceso, intermedias en la cadena causal entre exposición al tratamiento y efecto, **surrogate**). Esto se puede hacer con buenos y malos fines. Los buenos son intentar disponer de resultados más tempranamente, cuando los efectos últimos esperados del tratamiento tardan mucho tiempo en aparecer. Sólo se han de utilizar variables intermedias cuando hay pruebas empíricas de que son **predictores** adecuados **de las variables de resultado finales**. El problema que surge con estas variables intermedias a veces es que se desco-

noce si son buenos predictores de dichas variables finales o porque su capacidad predictiva y discriminante se ha estudiado de forma deficiente. (Este tema se describe en el Capítulo 6.) El artículo [Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being missed? Ann Intern Med. 1996;125:605-13](#) incluye ejemplos y esquemas muy ilustrativos de los mecanismos de las variables sustitutivas.

Por último, cabe destacar las denominadas **variables compuestas** (*composite variables*). Su uso en ensayos va en aumento hace tiempo. Cuando en los artículos o informes se incluyen, además, los resultados de las variables individuales que las integran, no hay problema. Pero este aparece cuando en los ensayos publicados solo se notifican las variables compuestas no desglosadas, es decir, se omiten las variables que las componen. El artículo [Montonri VM, et al Validity of composite end points in clinical trials. BMJ. 2005;330:594-6](#) ofrece una buena guía para revisar ensayos con estas variables y explica muy brevemente y con sencillez un ejemplo (tabla 1) en el cual los resultados no son concluyentes respecto a la mortalidad y al infarto de miocardio no mortal (los IC95% de los *hazard ratios* incluyen el 1) de dos tratamientos de la angina.

Al analizar en la tabla 1 el IC95% de las admisiones por síndrome coronario agudo, sí se observa que el tratamiento invasivo se asocia con una reducción del odds ratio (OR, ver Capítulo 4) de admisiones entre 0,21 y 0,45. El riesgo de muerte puede ser mayor con el tratamiento invasivo, pues oscila entre una reducción de riesgo de 0,72 y un aumento notable, de 3,16. Sin embargo, según la variable compuesta, el tratamiento invasivo reduce de 0,21 a 0,45 el riesgo de “lo que mida” la variable compuesta, algo que desconocemos, aunque es probable que sea el resultado de una combinación aritmética de algún tipo de las otras tres variables. Con esta información no hay problema al interpretar los resultados, pero, como se ha subrayado en numerosas publicaciones, el problema aparece cuando sólo se notifican los resultados de la variable compuesta sin desglosarla. Es fácil imaginar las conclusiones erróneas que se extraerían en este caso si se desconocieran los *hazard ratios* de las variables que se incluyen en la compuesta.

9. Análisis por intención de tratar

Los resultados obtenidos en cada participante en un ensayo pueden analizarse:

- 1) atribuyéndolos al tratamiento que realmente recibió el paciente (análisis por protocolo, *per protocol*, APP) al margen de que, de *facto*, algunos recibieran el tratamiento de otro grupo del ensayo (por ej., *switches*), distinto al que les correspondía según la asignación aleatoria o
- 2) atribuyéndolos al tratamiento del grupo al que fue aleatorizado con independencia del grado de cumplimiento y de que recibieran un tratamiento del otro grupo del ensayo (AIT, *intention-to-treat analysis*).

El método recomendado es el AIT, por varios motivos, fundamentalmente los siguientes. Si no se respeta el AIT, se viola el principio de aleatorización (las variables son aleatorias si se respeta su distribución conforme a la asignación aleatoria). Además, con el AIT se reduce la posible sobrestimación de la eficacia del tratamiento experimental, dado que en la práctica clínica habitual la efectividad observada suele ser menor que la eficacia estimada en las condiciones experimentales y con el alto grado de control de un ensayo clínico, mayor que en la práctica. Y, no menos importante, el AIT se considera la forma más adecuada de tratar los **errores de aleatorización**: incluir los pacientes que van a ser aleatorizados sobre la base de información basal incorrecta, aleatorización de pacientes que incumplen los criterios de selección, varias aleatorizaciones de un mismo paciente por error, pacientes aleatorizados que por error no reciben el tratamiento al que han sido asignados.

Cuando se comparan los resultados que se obtienen con ambos métodos, la eficacia estimada con el APP tiende a ser mayor (posible sobreestimación) que la estimada con el AIT. No obstante, como hay controversias sobre el método de elección, algunos recomiendan analizar los resultados por protocolo y por intención de tratar, comparar sus resultados y notificarlos. También se recomienda, teniendo en cuenta la subnotificación de los errores de aleatorización, notificarlos cuando se produzcan y no modificarlos para preservar el principio del AIT.

10. Métodos y mediciones (basal y sucesivas)

En los ensayos se emplean numerosos modelos y pruebas de análisis estadísticos. Es necesario revisarlos para evaluar si se utilizan los apropiados para los objetivos, diseños y variables empleados en el ensayo, si su aplicación es adecuada, y si se interpretan correctamente o pedir ayuda a quien tenga conocimientos avanzados en estadística.

También es esencial evaluar la adecuación de las mediciones que se realizan y cómo y cuándo se realizan, un tema en el cual el razonamiento y conocimiento clínico es importante. Las mediciones basales, justo antes de iniciar la administración de los tratamientos, antes de intervenir, informan sobre 1) la distribución y el valor de las variables en los grupos antes de la intervención, 2) por ende, la similitud de los grupos o la distribución desigual de alguna variable potencialmente asociada con los efectos (resultados) y, por ello, confusora (se explica en el Capítulo 5), y 3) los valores de referencia iniciales, sin intervenir (basales) con los cuales se compararán (pre-post intervención) los resultados finales del ensayo para estimar efectos causales (la eficacia y la seguridad de los tratamientos administrados o las intervenciones realizadas) en relación con el grupo control. Por último, debe prestarse atención a los análisis intermedios (*interim analyses*).

11. Resultados

Todo el conocimiento en clínica, epidemiología y estadística ha de aplicarse a la evaluación de los resultados de los ensayos clínicos. Este capítulo introductorio se limita a desta-

car varios hechos importantes en los ensayos clínicos, que a veces se desconocen, pasan desapercibidos o no se les concede la importancia que tienen.

11.1. Unidad de análisis, medias y varianzas

Es crucial tener claro que **en los ensayos clínicos** los efectos de los medicamentos o intervenciones que **se estiman** son **efectos medios**, no efectos a nivel individual. En ellos se estima, por ejemplo, qué porcentaje de pacientes sobreviven determinado tiempo (un año) en cada grupo tras la administración de un quimioterápico, el de los que responden y no responden a los tratamientos (conforme a las variables de resultado escogidas) en el grupo experimental y en el control. Sin embargo, todavía no es posible saber para todos los medicamentos qué pacientes concretos del grupo al que pertenecen (sea el experimental o el de comparación) responderán y cómo al tratamiento, en quiénes será o no tóxico, aquellos en que la enfermedad progresará o no. Hay quien sostiene que el ensayo ideal, disponiendo de los participantes necesarios en cada grupo (n_i), debería consistir en una combinación de n_i ensayos n -de-1, porque en ellos se estima el efecto a nivel de cada participante y puede entonces estimarse la variabilidad entre participantes, una información crucial. La medicina predictiva, estratificada y personalizada ha avanzado en este sentido con el desarrollo de biomarcadores de respuesta, pronóstico, toxicidad y otros predictivos, aunque menos de lo esperado, porque además de las estimaciones en subgrupos, todavía, en numerosas ocasiones, no se conoce su capacidad predictiva a nivel individual o esta es baja, un tema que se explicará en el Capítulo 6. En este hecho, que ha estimulado la innovación en nuevos diseños, como los paraguas o cesta comentados antes, radica parte de la importancia de los análisis de subgrupos.

11.2. Análisis de subgrupos

No es infrecuente que en los grupos de tratamiento de un ensayo se detecte heterogeneidad en las respuestas de diferentes subgrupos de pacientes (edad, sexo, grado de cumplimiento, raza, comorbilidad, etc.). Esto ocurre, por ejemplo, en un estudio que estima la supervivencia global estratificada en diferentes subgrupos (edad, sexo, raza, estado PD-L1 y BRAF, inmunoterapia previa, etc.) de los pacientes de un ensayo fase III con pembrolizumab frente a ipilimumab para el tratamiento del melanoma avanzado. La presentación exhaustiva y correcta de estos análisis se encuentra en la figura 2 de este artículo: [Robert C, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. N Engl J Med. 2015;372:2521-32.](#)

Dada la tendencia a sobrevalorar los resultados (sobreestimar efectos), el análisis de los resultados por subgrupos tiene que ser cuidadoso. Se ha comprobado que hay estudios que notifican resultados de subgrupos y omiten los globales porque los primeros son más favorables al nuevo tratamiento o a la nueva intervención. Sin disponer de ambos resultados, cualquier conclusión o acción que se emprenda es arriesgada. Ha de proce-

derse con mucha cautela al interpretar los resultados y al extraer conclusiones cuando los análisis de subgrupos no se han diseñado y planificado *a priori* y cuando no constan en el protocolo del ensayo (y, más aún, si esto no se indica en los artículos publicados). Si los cálculos del tamaño muestral *a priori* se han realizado para el conjunto de participantes para una potencia estadística determinada y no para subgrupos, al realizarlos en un subgrupo, como éste tiene un tamaño mucho menor que el conjunto, la potencia estadística decrece tanto más cuanto menor sea el tamaño del subgrupo (lo cual puede aumentar los errores de tipo II).

Los análisis por subgrupos siempre hay que valorarlos teniendo en cuenta sobre todo los no previstos en la fase de diseño del ensayo, porque sus resultados pueden estar sobreestimando, como se ha indicado, la eficacia o la seguridad del nuevo medicamento. En el artículo [Sun X, Briel M, Walter SD, Guyat GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340:c117](#) se incluyen las preguntas esenciales que deben responderse siempre al evaluar resultados en subgrupos.

11.3. La tiranía de las medias

Se olvida o desconoce que la imposición de valores medios estimados en grupos o poblaciones a los individuos de dichos grupos y poblaciones es, lamentablemente, frecuente en epidemiología. Este error crucial se conoce como la tiranía de las medias (*The tyranny of the means*).

La clave de este problema estriba en que las medidas “promedio” estimadas en grupos omiten, no pueden capturar, la variabilidad o heterogeneidad de y entre individuos de los grupos y las poblaciones, sin cuyo conocimiento no es posible discriminar resultados (quién desarrolla la enfermedad o el desenlace y quién no tras la exposición a nivel individual). Esto ocurre en los ensayos clínicos. **Los efectos que se estiman en ellos, en ausencia de análisis de subgrupos y sin ensayos n-de-1, son efectos medios en el grupo, no a nivel de cada integrante de los grupos experimental y control.** Por ejemplo, si en un ensayo clínico con un nuevo tratamiento de un tumor la supervivencia media estimada al año en el grupo experimental es del 45%, esto significa que, de cada 100 tratados, como media 45 sobrevivirán un año, pero no dice nada sobre lo que, a nivel individual, sobrevivirá cada paciente de ese grupo; unos sobrevivirán más, otros 1 año y otros menos de 1 año y 45, el promedio, 1 año, lo que siempre se comprueba en la práctica.

Si se consiguieran sustituir los ensayos tradicionales, por ejemplo, con un grupo experimental y un control, por tantos ensayos n-de-1 como participantes tuviera ese ensayo, podríamos tener estimaciones de los efectos de los tratamientos a nivel individual, de la variabilidad (heterogeneidad) entre individuos, de efectos en subgrupos y de efectos globales, el ideal. No obstante, en la actualidad hay métodos estadísticos (se llaman *jerárquicos*, *hierarchical*) con los cuales se pueden estimar efectos a estos tres niveles y obtener la máxima información necesaria, que, a la postre, es la que interesa.

12. Fases de un ensayo

Los ensayos se dividen en 4 fases, aunque hay variaciones en torno a ellas. En general:

Fase I. El tratamiento nuevo (experimental) se administra a un grupo reducido de personas sanas (10-40) para evaluar su seguridad, la farmacocinética y la dosificación.

Fase II. Se administra a un grupo mayor (de decenas a pocas centenas) de pacientes y, a veces, a un grupo control, para iniciar los estudios de eficacia y ampliar los de seguridad (toxicidad). Está aumentando la frecuencia de solicitudes de autorización de medicamentos con resultados promisorios en esta fase (temprana), que algunas agencias reguladoras conceden de forma condicional a la espera de que se recabe más información bajo determinadas condiciones, incluida su financiación, sobre su efectividad y seguridad.

Fase III. Es el ensayo que acostumbramos ver publicado con un grupo experimental y uno o más controles, de mayor tamaño (500-2000) y en él se realizan las estimaciones de eficacia y seguridad que se presentan para solicitar la autorización de medicamento.

Fase IV. Son los ensayos que se llevan a cabo tras la aprobación y comercialización del nuevo medicamento con miras a estimar su efectividad (que no eficacia) y mejorar las estimaciones de seguridad, habida cuenta de que pueden observarse respuestas diferentes y detectarse efectos secundarios porque se administra a muchos más pacientes (si el NNT es alto) de los que participaron en el ensayo en fase III y en condiciones de práctica clínica habituales, no experimentales y controladas.

En fase IV pueden utilizarse, y de hecho se utilizan, estudios observacionales con el mismo objetivo: recabar información adicional sobre efectividad y seguridad y **reducir**, así, la **incertidumbre asociada con las estimaciones de la eficacia y seguridad de los ensayos, siempre presente cuando en condiciones de práctica clínica los tratamientos se administran durante más tiempo que en el ensayo y a más pacientes y sin el grado de vigilancia y control que en los ensayos**. Por ejemplo, si el número necesario de pacientes que hay que tratar para observar un efecto secundario (*number needed to hurt* o NNH) es, afortunadamente, bajo, los efectos secundarios no se observarán en el ensayo sino cuando el medicamento se administre a muchos más pacientes tras su autorización y comercialización.

Por otro lado, una vez comercializado, un medicamento puede administrarse con indicaciones fuera de ficha técnica (*off label*) más o menos distintas de las del ensayo (efecto de extensión del tratamiento) y con mayor, menor o nulo control y vigilancia, a pesar de las normativas, lo cual aumenta más la incertidumbre sobre su efectividad y seguridad y el riesgo de que aparezcan efectos secundarios no detectados en fases anteriores de los ensayos.

En muchos de los hechos hasta aquí explicados estriba la importancia de la evaluación y de la vigilancia postautorización de la efectividad y la seguridad. Los estudios observa-

cionales de **efectividad comparada** (*comparative effectiveness*) con **datos de la vida real** (*Real World Data*) y la farmacovigilancia persiguen este fin: recabar información adicional sobre efectividad y seguridad en condiciones habituales de práctica clínica (no experimentales ni controladas) y compensar así el margen de incertidumbre siempre asociado con los resultados de los ensayos clínicos, la brecha entre eficacia (efecto estimado en los ensayos) y efectividad (efecto evaluado en condiciones de práctica clínica habitual).

Aunque también trascienda los objetivos de este cuaderno, debe apuntarse que con los llamados **ensayos pragmáticos** (*pragmatic trials*) también se pretende reducir dicha brecha, la sobrestimación de la eficacia y subestimación de la seguridad de los ensayos clínicos convencionales, así como aumentar su validez externa, aunque a expensas de disminuir algo su validez interna. Estos ensayos, a diferencia de los ensayos convencionales o explicativos (*explanatory*), se denominan *pragmáticos* porque con ellos, al intentar remedar mucho más de cerca las condiciones de la práctica clínica real, se persigue como objetivo importante, además de estimar *efectividad* y seguridad, informar decisiones clínicas y de política sanitaria aportando información obtenida en esas condiciones reales mucho más cercanas a la práctica de la medicina, no experimentales.

13. Consideraciones finales

Entre las numerosas consideraciones que podrían hacerse sobre el amplio abanico temático que abarcan los ensayos clínicos, cabe subrayar las siguientes de forma muy sucinta, por no haber sido tratadas hasta ahora.

Es necesario distinguir entre **significación estadística** y **significación clínica**, porque no siempre van a la par, sobre todo cuando las muestras son grandes. Un resultado estadísticamente muy significativo puede ser clínicamente irrelevante. No olvidemos que siempre puede calcularse cuál es el tamaño muestral necesario para que una diferencia entre porcentajes o medias entre grupos o cualquier otro parámetro sea estadísticamente significativa por pequeña en magnitud e irrelevante que sea desde el punto de vista clínico.

La **repetibilidad** de los ensayos es baja cuando se considera la población de todos los ensayos realizados, a juzgar por la bajísima fracción de ensayos que se llevan a cabo con los mismos objetivos y tratamientos o que se repiten. Las revisiones y metanálisis que se realizan con ellos ponen de relieve que a veces sus resultados son no concluyentes o son contrapuestos. De ahí la necesidad de ser extremadamente cautelosos al tomar decisiones sobre los resultados de un único ensayo.

No debe olvidarse que cuando se usa **placebo** en un ensayo, unos pacientes responden a él y otros no y que tampoco es posible predecir estas respuestas a escala individual.

Siempre hay que revisar la “**Tabla 1**”. La asignación aleatoria no siempre logra la distribución homogénea de todas las variables potencialmente asociadas con el resultado entre los grupos del ensayo clínico. Además de valorar la relevancia y la justificación clínicas de las

variables utilizadas en él y las variables importantes que se omitieron, así como evaluar sus efectos en la interpretación de los resultados (sesgos, confusión, sobre o infraestimación de eficacia y seguridad), es preciso ver si se realizaron ajustes *a posteriori* de las medidas de resultados, de los estimadores de los efectos, como tasas de mortalidad, de supervivencia, de recidiva, RR, OR o HR, de las variables de resultado cuyas distribuciones aleatorias fueron marcadamente distintas tras la asignación aleatoria.

Aunque las autoridades regulatorias (EMA, FDA) instan a estimar la **eficacia y la seguridad relativas**, el porcentaje de ensayos clínicos realizados que lo hacen aún no supera, como mucho, el 50% según sus propias estimaciones. A ello cabe añadir que, entre los que las estiman, la fracción de aquellos con eficacias relativas marginales (muy bajas) es elevada cuando se considera toda la población de ensayos realizados anualmente.

Recuérdese asimismo que el diseño experimental puede aplicarse en numerosos **ámbitos** ajenos al medicamento y a la clínica.

Hay **estudios observacionales** con los cuales se pueden obtener, si se cumplen ciertos requisitos, grupos de estudio o intervención y grupos control similares a los que se obtienen por medio de la asignación aleatoria, aunque los resultados, al estar más expuestos a amenazas a la validez interna, han de interpretarse con sumo cuidado. Se explican en el Capítulo 7.

Cabe hacer hincapié en la importancia de la farmacovigilancia y de los estudios (observacionales mencionados en el párrafo anterior) de **efectividad comparada** (*Comparative effectiveness*) con datos de vida real. A semejanza de los ensayos pragmáticos, con estos estudios se obtiene información que reduce la brecha entre eficacia y efectividad y pueden ofrecer información clave sobre seguridad en condiciones de práctica clínica habitual, no experimentales (ver el capítulo 7).

14. Material didáctico complementario

Stupp R, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* 2009;10:459–66.

Harrison CN, et al., Ruxolitinib versus Best Available Therapy for ET intolerant or resistant to Hydroxycarbamide in a Randomized trial. *Blood.* 2017;26;30:1889-97.

Fojo T, Mailankody S, Andrew Lo A. Unintended Consequences of Expensive Cancer Therapeutics—The Pursuit of Marginal Indications and a Me-Too Mentality That Stifles Innovation and Creativity The John Conley Lecture. *JAMA Otolaryngology–Head Neck Surg.* 2014;140:1225-36.

Freidlin B, et al. Two-by-Two Factorial Cancer Treatment Trials: Is Sufficient Attention Being Paid to Possible Interactions? *JNCI.* 2017;109. doi: 10.1093/jnci/djx146.

Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. NEJM.2017;377:62-70.

Park JW, et al. Adaptive Randomization of Neratinib in Early Breast Cancer N Engl J Med 2016;375:11-22.

Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being missed? Ann Intern Med. 1996;125:605-13.

Montori VM, et al Validity of composite end points in clinical trials. BMJ. 2005;330:594-6

Robert C, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. N Engl J Med. 2015;372:2521-32.

Sun X, Briel M, Walter SD, Guyat GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340:c117.

15. Bibliografía

General

Nivel introductorio

- Gordis L. Epidemiology, Second Edition. Philadelphia, PA: WB Saunders Company; 2000.
- Mausner JS, Kramer S. Epidemiology: An introductory text. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. Foundations of epidemiology. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epiemiología clínica. Ciencia básica para la medicina. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. Epidemiology, Beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc.; 2000.

Específica de este Capítulo

- En las referencias del nivel introductorio y del intermedio hay capítulos que tratan estos temas en detalle, con numerosos ejemplos y mucha claridad.
- Meinert CL. Clinical trials: design, conduct and analysis. New York: Oxford University Press; 1986.

- CONSORT Statement and guidelines for clinical trials: <http://www.consort-statement.org/>
- CONSORT extensions for: cluster trials, Non-inferiority and equivalence trials, Pragmatic trials, N-of-1-trials, Pilot and feasibility trials: <http://www.consort-statement.org/extensions>
- Extension for harms (safety): <http://www.consort-statement.org/extensions?ContentWidgetId=561>
- Fleming TR, DeMets DL. Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med.* 1996;125:605-613.
- Venook AP, Tabernero J. Progression-free survival: Helpful biomarker or clinically meaningless end point? *J Clin Oncol.* 2015;33:4-7.
- Ford I, Norrie J. Pragmatic trials. *N Engl J Med.* 2016;375:454-63.
- Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of Multi-Arm Parallel-Group Randomized Trials Extension of the CONSORT 2010 Statement. *JAMA.* 2019;321:1610-20.

Capítulo 4.

Estudios observacionales, medidas de asociación y sesgos

Índice

1. Estudios observacionales: clasificación	56
2. Estudios de cohortes.....	56
2.1. Medidas de asociación en los estudios de cohortes y ejemplos	57
3. Estudios de casos y controles	59
3.1. Medidas de asociación en los estudios de casos y controles y ejemplos .	61
4. Estudios transversales o de prevalencia	65
4.1. Medidas de asociación en los estudios transversales y ejemplos	66
4.2. Prevalencia y relación entre medidas de asociación	66
5. Sesgos	67
6. Otras aplicaciones de los estudios observacionales	71
7. Bibliografía	72

Objetivos específicos de este Capítulo

1. Conocer las principales características del diseño, la realización, el análisis y la interpretación de resultados de los estudios observacionales de cohortes, casos y controles y transversales.
2. Conocer las diferencias entre estudios experimentales y observacionales.
3. Conocer los principales tipos de sesgos que pueden introducirse en estos estudios, así como sus consecuencias en los resultados.

Nota importante

En este Cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I, 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como **Material didáctico complementario**.

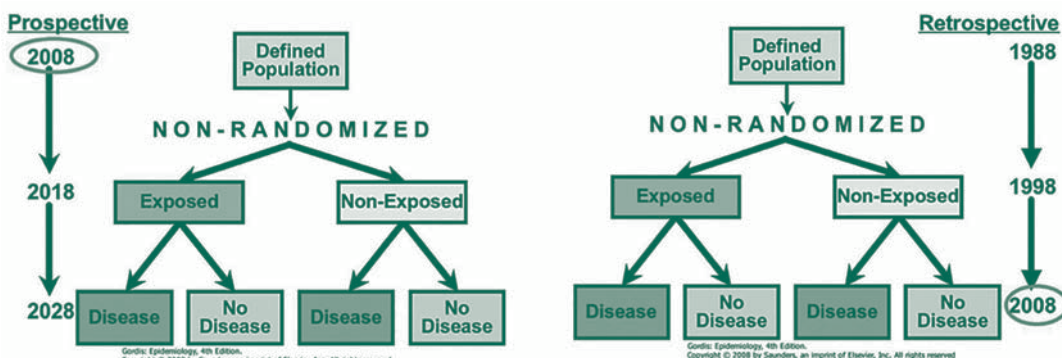
1. Estudios observacionales: clasificación

Antes de empezar, recomendamos releer el apartado 3 (Tipos de estudios) del Capítulo anterior. Las relecturas ayudan a asentar los conceptos. En este Capítulo pasaremos a describir acto seguido los estudios observacionales (o no experimentales), concretamente los de cohortes, los de casos y controles y los transversales. En el Capítulo 7 se describen otros estudios observacionales. No se hace aquí porque para entenderlos se necesita la información de los Capítulos 5 y 6. Estos estudios son las variables instrumentales, los estudios de diferencias en diferencias, las series temporales truncadas con regresiones segmentadas. Además, estudiaremos un método de emparejamiento de los participantes en dos grupos de un estudio, el propensity score.

De entrada, es sumamente importante recordar siempre que la ausencia de asignación aleatoria es la principal limitación de los estudios observacionales, porque sin ella no es nada fácil disponer de grupos de estudio (grupo experimental o de intervención y grupo control) en los cuales la distribución de las *variables observadas (observables)* y *no observadas (unobservables)*, porque no se hayan medido o porque se desconozcan) asociadas con los resultados sea similar y permita, entonces, hacer comparaciones entre los grupos para estimar los efectos de las intervenciones (efectos causales) en las cuales no intervengan sesgos ni factores de confusión, dos conceptos esenciales que se explicarán en este y en el siguiente capítulo.

2. Estudios de cohortes

Los estudios de cohortes, aunque siempre se dirigen de la exposición a la aparición de un resultado, pueden realizarse, como ilustran las siguientes figuras, de forma concurrente (izquierda) y no concurrente (derecha). En la primera no se dispone de información sobre los resultados al iniciar el estudio y la segunda puede llevarse a cabo cuando sí contamos con la información necesaria (en registros) sobre la exposición y el desenlace.



Estos estudios son costosos (mucho más que los de casos y controles y los transversales) y por ello conviene realizarlos cuando ya disponemos de pruebas que sugieran la existencia de una asociación entre exposición y desenlace. Sin embargo, hay que tener en cuenta que su viabilidad depende también de que, primero, el tiempo que transcurra entre exposición y desarrollo de la enfermedad (o de cualquier otro desenlace) no sea demasiado prolongado, segundo, se espere una baja tasa de abandonos o pérdidas en el seguimiento (*attrition*), tercero, se pueda encontrar un grupo de expuestos y no expuestos (que realmente no tengan la enfermedad o desenlace) y, cuarto, la incidencia del desenlace no sea demasiado baja. Estos estudios complejos y costosos no se llevan a cabo si hay que esperar muchos años para observar los desenlaces o si cabe esperar que aparezcan muy pocos casos por tratarse de una enfermedad o de cualquier otro resultado con muy baja incidencia (probabilidad de aparición).

No obstante, en este último caso, si se dispone de la información necesaria, es posible realizar un estudio de cohortes no concurrente, como se esquematiza a continuación:



2.1.Medidas de asociación en estudios de cohortes y ejemplos

La medida de asociación en los estudios de cohortes es el riesgo relativo (*relative risk*). Para calcularla, primero se identifican los expuestos y no expuestos, a continuación, los que desarrollan y no desarrollan la enfermedad en ambos grupos, y, por último, se calcula la incidencia en expuestos y no expuestos como muestra el siguiente esquema mediante una tabla 2 x 2.

		Disease develops	Disease does not develop	Totals	Incidence of disease
First, identify	Exposed	a	b	a + b	$\frac{a}{a + b}$
	Not exposed	c	d	c + d	$\frac{c}{c + d}$

$$\frac{a}{a + b} = \text{Incidence in exposed} \quad \frac{c}{c + d} = \text{Incidence in not exposed}$$

Aquí se presenta otro ejemplo para estimar la incidencia de enfermedad coronaria (coronary heart disease, CHD) en fumadores (expuestos) y no fumadores (no expuestos).

First,
select

	Develop CHD	Do not develop CHD	Totals	Incidence of disease
Smoke cigarettes	84	2916	3000	$\frac{84}{3000}$
Do not smoke cigarettes	87	4913	5000	$\frac{87}{5000}$

$$\frac{84}{3000} = 0.028 = \text{Incidence in 'smoke cigarettes'}$$

$$\frac{87}{5000} = 0.0174 = \text{Incidence in 'not smoke cigarettes'}$$

El denominado **riesgo absoluto** (o incidencia) traduce la magnitud del riesgo de enfermedad de un grupo o población con cierta exposición. Como no incorpora el riesgo en los que no están expuestos, no indica si la exposición está asociada o no con la enfermedad. Para averiguar si existe dicha asociación, debe comprobarse que el riesgo es mayor en los expuestos (e) que en los no expuestos (ê). A partir de la tabla anterior podemos calcular:

La diferencia de riesgos = (incidencia en expuestos) – (incidencia en no expuestos) =

$$= 0,028 - 0,0174 = 0,0106$$

y **la razón de riesgos** (o de incidencias) o **riesgo relativo** (RR) =

(incidencia en expuestos) / (incidencia en no expuestos) =

$$= 0,028 / 0,0174 = 1,61$$

El RR es la medida de asociación de los estudios de cohortes y los ensayos clínicos, es decir de los estudios longitudinales prospectivos. Sólo se puede calcular cuando partimos de la exposición y nos dirigimos en el tiempo hacia delante, hacia el desenlace, como se ha indicado anteriormente, porque así se pueden detectar durante el seguimiento los **casos nuevos o incidentes**.

Es importante apreciar las diferencias entre los tres conceptos: 1) riesgo absoluto, 2) diferencia de riesgos y 3) riesgo relativo porque indican cosas distintas y se confunden con frecuencia: 1) magnitud del riesgo, 2) exceso de riesgo en los expuestos, y 3) magnitud del riesgo en los expuestos respecto a los no expuestos, respectivamente. Veamos otro ejemplo para ilustrarlo.

	Grupo A	Grupo B
Incidencia en expuestos (I_e)	40	90
Incidencia en no expuestos ($I_{\bar{e}}$)	10	60
Riesgo absoluto = $I_e - I_{\bar{e}}$	30	30
$RR = I_e / I_{\bar{e}}$	4,0	1,5

El riesgo absoluto (en expuestos) es mayor en B (90) que en A (40), la diferencia de riesgos es la misma (30) y el RR es más del doble en A (4,0) que en B (1,5).

Cuando la incidencia es la misma en los expuestos y en los no expuestos, el $RR = 1$, y no hay asociación entre la exposición y la enfermedad.

Si el $RR > 1$, el riesgo en los expuestos es mayor que en los no expuestos y existe esa asociación (que puede ser causal o no, como se explicará en el Capítulo 7).

Si el $RR < 1$, ocurre lo contrario que en el caso anterior: el riesgo en los expuestos es menor que en los no expuestos y la exposición protege del desarrollo del efecto. En este caso también se deberá valorar si esa asociación es causal y, por lo tanto, protectora de desarrollar la enfermedad.

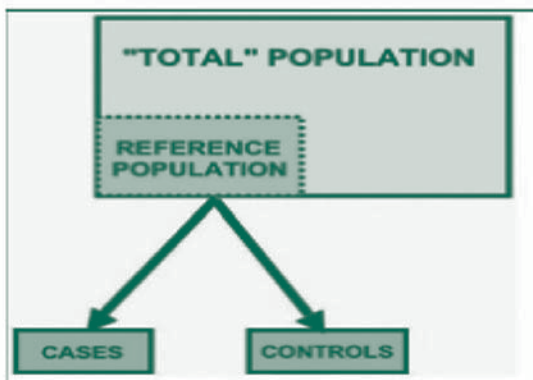
Todas estas estimaciones deben acompañarse de sus intervalos de confianza para conocer la incertidumbre asociada con ellas.

Insistamos en que, al tratarse de estudios no aleatorizados, cuando con estos estudios se encuentran asociaciones deben tomarse precauciones respecto a los sesgos y las variables de confusión. Los sesgos se verán en este Capítulo, un poco más abajo, y la confusión, en el 5 y el 7.

3. Estudios de casos y controles

Como ilustran los siguientes esquemas, en los estudios de casos y controles, se procede en sentido contrario al de los de cohortes: primero se seleccionan los enfermos y no enfermos (casos y controles), a continuación se observa cuántos de ellos habían estado expuestos y no expuestos y, seguidamente, se estima el porcentaje de expuestos y no expuestos en cada uno de ambos grupos. A diferencia de los estudios de cohortes, en los cuales los casos son nuevos o incidentes (van apareciendo durante su realización, son prospectivos), en los de casos y controles, no lo son (aparecieron antes del inicio del estudio) y, por consiguiente, no se pueden calcular incidencias; la enfermedad en los casos puede haberse desarrollado hace tiempo. Por esta razón, en los de cohortes la medida de la fuerza de la asociación que se estima es el riesgo relativo (incidencia en expuestos y en no expuestos) y en los de casos y controles, la odds ratio o razón de productos cruzados, como se explica más adelante.

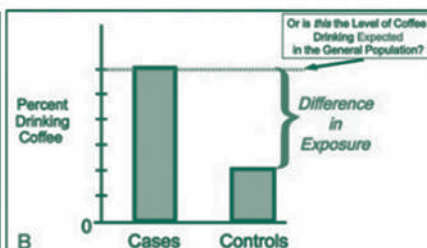
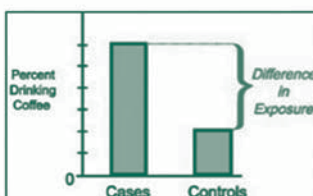
En los estudios de casos y controles, los casos han de ser representativos del conjunto de personas con la enfermedad o la característica en estudio en la población general, y los controles, de las personas sin la enfermedad en la población de la cual proceden los casos, como se muestra en esta figura:



<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Los controles han de ser lo más similares posible a los casos excepto en que no tengan la enfermedad. Además, **han de seleccionarse de tal forma que dicha selección sea independiente de la exposición, en el sentido de que su selección no sea más o menos probable si están expuestos**. La finalidad última es que el porcentaje de expuestos y no expuestos sea lo más similar posible al de la población general de la cual proceden, como ilustran las siguientes figuras con las preguntas que en ellas se formulan, porque, de lo contrario, la sobre o infrarrepresentación de expuestos y no expuestos desemboca en sobre o infraestimación de la magnitud real de la OR. Recuérdese siempre que **las sobre-representaciones de expuestos y no expuestos en casos y controles sesgan las estimaciones de las medidas de asociación (hacia la sobre o subestimación), algo que debe evitarse a toda costa**.

Did patients with **cancer** of the pancreas drink **more coffee** than did people without cancer of the pancreas in the **same population**?



Gordis, Epidemiology, 2013

Los criterios de inclusión y exclusión utilizados en la selección de los casos que no estén relacionados directamente con la enfermedad o característica en estudio deben aplicarse también con rigor a los controles (por ejemplo, no padecer diabetes o tener entre 20 y 70 años de edad).

Insistamos en que, al tratarse de estudios no aleatorizados, cuando con estos estudios se encuentran asociaciones deben tomarse las precauciones respecto a los sesgos y las variables de confusión. Los sesgos se verán en este Capítulo, un poco más abajo, y la confusión, en el 5.

3.1. Medidas de asociación en los estudios de casos y controles y ejemplos

La forma de proceder para calcular la **odds ratio** (OR), la medida de asociación en los estudios de casos y controles, es la siguiente: primero, se seleccionan los casos y los controles y a continuación se averigua cuáles estuvieron expuestos y no expuestos:

		First, select	
		Cases (with disease)	Controls (without disease)
Then, measure past exposure	Were exposed	a	b
	Were not exposed	c	d
	Totals	a + c	b + d
Proportion exposed		$\frac{a}{a + c}$	$\frac{b}{b + d}$

Ejemplo de cálculo en los estudios de casos y controles:

	Cases CHD	Controls (without disease)
Smoked cigarettes	112	176
Did not smoke cigarettes	88	224
Total	200	400
% Smoking cigarettes	$\frac{112}{200} = 56\%$	$\frac{176}{400} = 44\%$

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Acto seguido, se calcula el **odds**, que se define así:

$$\text{odds} = \frac{\text{probabilidad de que ocurra un evento}}{\text{probabilidad de que no ocurra}} = \frac{p}{(1-p)}$$

Por consiguiente, y esta distinción es muy importante, odds y probabilidades no son lo mismo, como se muestra en el siguiente ejemplo.

	Develop Disease	Do Not Develop Disease
Exposed	a	b
Non-exposed	c	d
The probability that an exposed person develops disease = $\frac{a}{a+b}$		
The probability that a non-exposed person develops disease = $\frac{c}{c+d}$		

	Develop Disease	Do Not Develop Disease
Exposed	a	b
Non-exposed	c	d
The odds that an exposed person develops disease = $\frac{a}{b}$		
The odds that a non-exposed person develops disease = $\frac{c}{d}$		

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

El OR se calcula entonces del siguiente modo y, por este motivo, también se conoce como razón de productos cruzados:

	Develop Disease	Do Not Develop Disease
Exposed	a	b
Non-exposed	c	d
Odds ratio is the ratio of the odds of disease in the exposed to the odds of disease in the non-exposed		

$$\text{OR} = \frac{\text{odds that an exposed person develops the disease}}{\text{odds that a non-exposed person develops the disease}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a}{b} \times \frac{d}{c} = \frac{ad}{bc}$$


<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

El OR se interpreta como el RR: OR = 1, no hay asociación entre exposición y enfermedad, puesto que el riesgo es idéntico en ambos grupos; si es > 1 el riesgo en los expuestos es mayor que en los no expuestos (hay asociación), y si es < 1, hay asociación y, si fuese causal, sería protectora de desarrollar la enfermedad, porque el riesgo en los expuestos es menor que en los no expuestos.

El OR puede ser un buen estimador del RR y puede estimarse en estudios prospectivos, pero sólo si se cumplen tres requisitos: 1) los casos han de ser representativos de los casos de la población general, 2) los controles han de ser representativos de los que no tienen la enfermedad en la población general, y 3) la prevalencia de la enfermedad ha de ser baja (la cifra se establece en < 1 / 10.000), como se aprecia seguidamente:

Comparing OR to RR: Disease Is NOT Infrequent

	Develop Disease	Do not Develop Disease	
Exposed	50	50	100
Non-Exposed	25	75	100

$$\text{Relative Risk} = \frac{50/75}{50/25} = 2$$

$$\text{Odds Ratio} = \frac{50 \times 75}{50 \times 25} = 3$$

Comparing OR to RR: Disease Is Infrequent

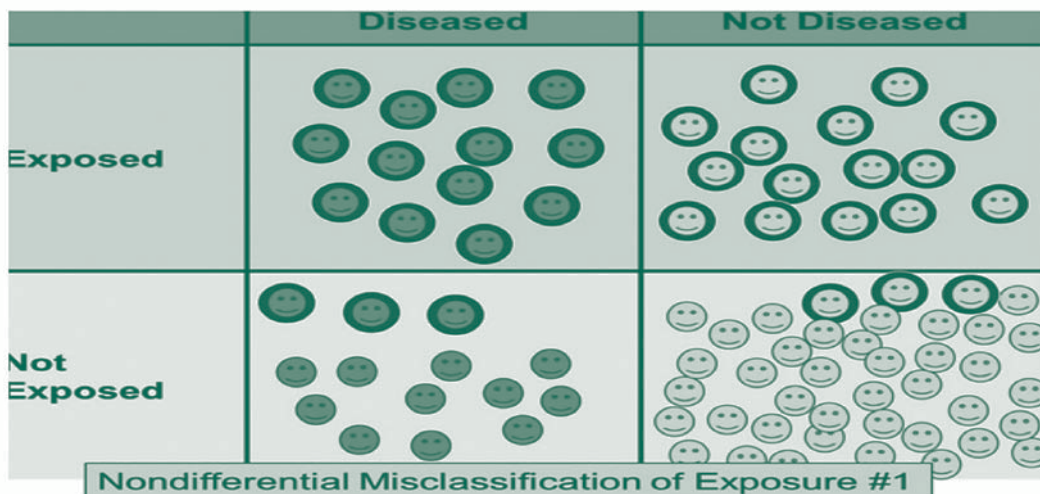
	Develop Disease	Do not Develop Disease	
Exposed	200	9800	10,000
Non-Exposed	100	9900	10,000

$$\text{Relative Risk} = \frac{200/10,000}{100/10,000} = 2$$

$$\text{Odds Ratio} = \frac{200 \times 9900}{100 \times 9800} = 2.02$$

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Los casos y los controles han de definirse con la mayor exactitud posible y ser representativos de la población de personas con y sin la enfermedad o el desenlace que se estudie, respectivamente. Es fundamental utilizar pruebas y criterios diagnósticos de alta validez para evitar clasificaciones erróneas que introduzcan sesgos de clasificación errónea (*misclassification bias*), que pueden desembocar en sub o sobrestimaciones de los RR y los OR por casos que en realidad son controles y viceversa, como ilustra la siguiente figura. En ella se ven clasificaciones erróneas de casos y controles verdaderos y realmente expuestos que se definen por error, respectivamente, como casos (3) y controles (3) no expuestos.



Como consecuencia, en este ejemplo se reducen la a y la b en la tabla 2 x 2 y aumentan (erróneamente) c y d. Dependiendo de la magnitud de cada uno de estos errores, cambiará la magnitud del OR resultante. Si, por ejemplo, los casos expuestos verdaderos se clasifican por error como casos no expuestos, se reduce a y aumenta c, con lo cual disminuye el

numerador del OR, aumenta el denominador, y el OR disminuye, lo que desemboca en una subestimación del OR verdadero.

Los controles deberían proceder de la misma población en riesgo de la cual provienen los casos (aunque en la práctica es frecuente que se seleccionen de otra población o subgrupo) y de forma que sean parecidos a los casos respecto a variables importantes, pero sin la enfermedad. No es fácil definir siempre con exactitud la población que está en riesgo. (Ver más adelante el sesgo de Bergson para más detalles.)

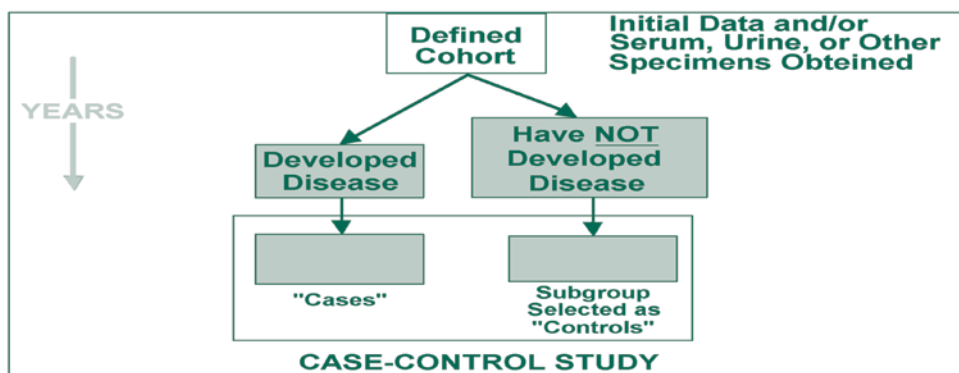
La selección de diferentes tipos de controles es posible, al igual que seleccionar más de un control por caso, pero en estas situaciones deben tomarse precauciones en el diseño, análisis e interpretación de los resultados, cuya explicación trasciende el alcance de este cuaderno introductorio.

El **emparejamiento** en los estudios de casos y controles (*matching*) consiste en la selección de controles que sean muy similares a los casos respecto a variables que puedan ser confusoras de las asociaciones que se encuentren (por ej., edad, comorbilidad, nivel de ingreso) y puede realizarse caso a caso (*matched pairs* o por parejas) o por grupos (*group matching*). En el Capítulo 5 se explica la confusión (*confounding*), que el emparejamiento es una de las medidas a nuestro alcance para evitarla, así como los efectos de los desequilibrios en la distribución de variables asociadas con (o pronósticas de) los resultados entre los grupos del estudio. Otra opción para emparejar es utilizar el denominado *propensity score*, que se explica en el Capítulo 7.

Es difícil encontrar controles adecuados cuando se intenta emparejar por más de dos variables (por ej., por edad y sexo y una tercera al mismo tiempo). Quede claro que al emparejar no se puede averiguar la existencia de asociaciones entre la enfermedad que se estudia y las variables por las cuales se han emparejado los casos y los controles, porque al formar pares no hay diferencias entre casos y controles respecto a estas variables, su distribución es la misma en ellos.

Un problema adicional es el emparejamiento no planificado, que se produce cuando al extraer casos y controles de una misma población, sobre todo cuando ésta es de tamaño reducido, más o menos casos y controles son similares por razones como la vecindad en relación con diversas variables, no de forma natural (cabe esperar, por ejemplo, que en un barrio de renta alta o baja los casos y controles que extraigamos serán similares en algunos hábitos y estilos de vida, educación). Al extraer casos y controles de poblaciones mucho mayores es posible reducir dicho emparejamiento.

Un diseño especial es el de **casos y controles anidado** (*nested case-control study*). Cuando durante el seguimiento de expuestos y no expuestos en un estudio de cohortes ya han aparecido casos (incidentes), es posible iniciar simultáneamente un estudio de casos y controles. Su diseño se esquematiza a continuación:



Design of a case-control study initiated within a cohort.

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Insistamos de nuevo en que, al tratarse de estudios no aleatorizados, cuando con estos estudios se encuentran asociaciones deben tomarse precauciones respecto a los sesgos y las variables de confusión. Los sesgos se verán en este Capítulo, un poco más abajo, y la confusión, en el 5 y el 7, y que lo mismo ocurre con los estudios de prevalencia que se explican a continuación.

4. Estudios transversales o de prevalencia

Estos estudios (*cross-sectional*) son observacionales y en ellos no se puede conocer la relación temporal entre exposición y enfermedad, porque la exposición (y no exposición) y la enfermedad (y la ausencia de ella) se consideran en el mismo momento en el tiempo en cada individuo.

Los **casos** considerados en ellos son **prevalentes** (no nuevos o incidentes, ya existen) y, por este motivo, pueden no ser representativos de todos los casos de enfermedad que se hayan desarrollado (por ejemplo, de los que fallecieron después de haber aparecido la enfermedad y antes de iniciar el estudio). Incluso si se detectara una asociación entre exposición y enfermedad, en estos estudios esta asociación se relaciona más con la supervivencia de los casos que con el riesgo de desarrollar la enfermedad (no son incidentes), un riesgo que no podemos conocer como en los estudios de cohortes o en los ensayos clínicos, porque no partimos de los expuestos y no expuestos y los seguimos en el tiempo para estimar las incidencias en ambos grupos, ni en los de casos y controles utilizando el OR como estimador del RR. Por otro lado, a veces es muy difícil saber si la exposición supuesta precede realmente a la enfermedad, un criterio de causalidad como se verá en el Capítulo 7.

En estos estudios se estiman las prevalencias de exposición y de enfermedad en poblaciones o grupos, que sirven, además de para conocer la magnitud de exposición y la carga de enfermedad con fines de planificación, para sugerir hipótesis sobre asociaciones causales (efectos causales, no para confirmar causalidad, como se verá en el Capítulo 7) que justifiquen la realización de alguno de los otros tipos de estudios que aportan más pruebas (evidencia) para estimar efectos causales de exposiciones. La **capacidad de los diferentes**

tipos de estudios para sugerir relaciones causales es, por orden decreciente, la siguiente: ensayos clínicos (estudios experimentales), estudios de cohortes, estudios de casos y controles, estudios de prevalencia, series de casos y opinión de expertos.

Cross-Sectional



<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

4.1. Medidas de asociación en los estudios transversales y ejemplos

Revisemos cómo se realizan los cálculos en los estudios de prevalencia:

	Disease	No Disease
Exposed	a	b
Not Exposed	c	d
Prevalence of exposure in Disease and No Disease		
	$\frac{a}{a+c}$	vs. $\frac{b}{b+d}$

	Disease	No Disease
Exposed	a	b
Not Exposed	c	d
Prevalence of disease in Exposed compared to Not Exposed		
	$\frac{a}{a+b}$	vs. $\frac{c}{c+d}$

4.2. Prevalencia y relación entre medidas de asociación

La razón de prevalencias (P) estimada en estos estudios sería =

$$P \text{ en expuestos} / P \text{ en no expuestos}$$

Hay una relación general entre prevalencia (P), incidencia (I) y duración de la enfermedad (D):

$$P = I \times D$$

que en estos estudios puede expresarse como razones entre expuestos y no expuestos:

$$\frac{Pe}{Pê} = \frac{Ie}{Îê} + \frac{De}{Dê}$$

Cuando D es muy corta (horas o muy pocos días, como en una diarrea bacteriana o vírica autolimitada o en la gripe), la duración en los expuestos (De) y no expuestos (Dê) es la misma y se anula la razón de duraciones, con lo cual, sólo en este caso, la razón de prevalencias puede ser un estimador de la de incidencias o RR:

$$\frac{Pe}{Pê} = \frac{Ie}{Îê}$$

5. Sesgos

Recordemos de nuevo que cuando se encuentra una asociación entre variables siempre hemos de averiguar si la explica el azar, si se debe a sesgos o a algún factor de confusión (estos últimos se explicarán en el Capítulo 5).

Los sesgos son **errores sistemáticos** (no aleatorios) en el diseño, la realización o el análisis de un estudio que producen estimaciones erróneas (sobre o infraestimaciones) del efecto que tiene una exposición (o del desenlace o resultado de que se trate) y de la fuerza de la asociación entre exposición y resultado.

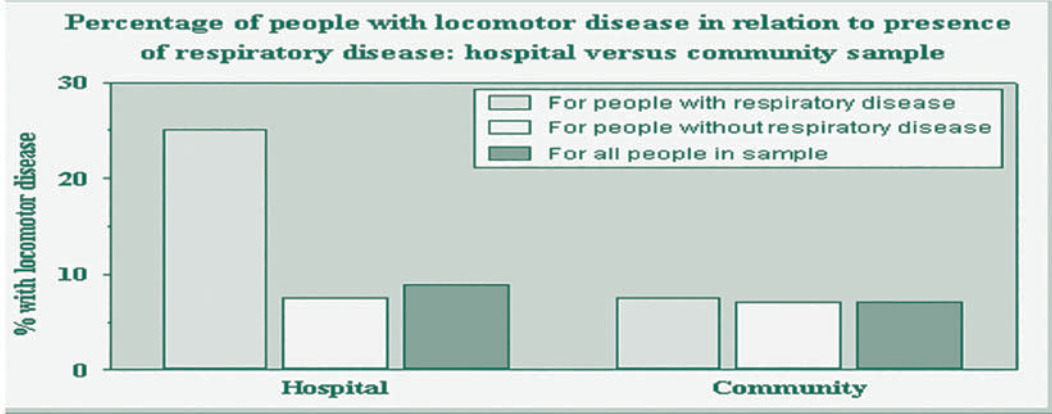
Al evaluar un estudio **no es aceptable desde el punto de vista científico limitarse a decir que hay un sesgo cuando este se detecta o sospecha. Hay que confirmar que existe, describir (no sólo nombrar) qué tipo de sesgo es y explicar qué efectos produce dicho sesgo en las estimaciones que se habían realizado**, es decir, mostrar por qué estas estimaciones son erróneas y en qué dirección las ha producido el sesgo: sobre o subestimación del efecto real (del RR, el OR o el HR o de la medida de efecto que se utilice).

Si bien hay muchos tipos de sesgos (la descripción de buena parte de ellos puede encontrarse en los manuales de epidemiología de la bibliografía de este módulo), cuatro son los tipos más importantes.

El **sesgo de selección** se produce cuando la forma de seleccionar casos y controles distorsiona (sobre o subestima) la magnitud real de una asociación entre exposición y desenlace (efecto) que existe en la población real, provocando que aparezca una **asociación espuria** o falsa. (No olvidemos que las asociaciones se estiman en muestras que proceden de una población y que lo que realmente nos interesa en última instancia no es lo que ocurre en la muestra sino en la población de origen. Los resultados observados en muestras se extrapolan a la población de la cual procede.)

Un ejemplo de este sesgo es el conocido como **sesgo de Bergson** (*Bergson's bias*), que puede introducirse cuando los casos y los controles no se extraen de la población de casos y controles general sino de una subpoblación, como puede ser la de un hospital. En el hospital es más probable encontrar pacientes con más de un factor de riesgo y más de una enfermedad que en la población general (en ella la mayor parte de las personas están sanas).

En la siguiente figura se muestra, a modo de ejemplo, que en las personas con una enfermedad del aparato locomotor el porcentaje de las que padecen una enfermedad respiratoria es mucho más alto en la muestra extraída del hospital que en la obtenida en la comunidad. Si el estudio se circunscribe al hospital, probablemente se concluya que hay una asociación entre ambas enfermedades y, además, que los factores de riesgo de la enfermedad respiratoria también lo son de la del aparato locomotor. En otras palabras, es probable que se encuentre una asociación inexistente, espuria, que no se encontraría con casos y controles extraídos de la población general y que es consecuencia de este sesgo de selección. Por ello es importante en situaciones como esta seleccionar casos de diferentes hospitales y, si es posible, que estos sean de distintos niveles.



Siempre que se encuentren diferencias entre los porcentajes de expuestos y no expuestos en casos y controles hemos de preguntarnos y, si es posible, confirmar, también, si la fracción de expuestos en los controles (la *b* en la casilla de la tabla 2 x 2) es realmente la esperada en la población general de la cual proceden los controles. Si es notablemente más alta o más baja que en la población de origen, la *b* de la tabla 2 x 2 será, respectivamente, mayor o menor que la real, y al calcular el OR ($a \times d / c \times b$) éste será subestimado (si es más alta) o sobreestimado (si es más baja); está en el denominador.

Otro ejemplo de este sesgo viene dado porque casos y controles también se seleccionan entre pacientes hospitalizados. Por ejemplo, como en el hospital es más probable encontrar personas que padecen alcoholismo y neumonía simultáneamente que aquellas que solo padecen neumonía, respecto a la población general, la fuerza de la asociación entre alcoholismo (exposición) y neumonía (efecto) puede sobreestimarse, como se ilustra a continuación (repárese en el paso de efectivos de *b* a *a* en las tablas 2 x 2; en el hospital los casos expuestos a alcoholismo (*a*) están sobreestimados respecto a la población general):

■ In the community		Pneumonia	
		Yes	No
Alcoholism	Yes	10	10
	No	90	90
OR = $\frac{10 \times 90}{90 \times 10} = 1.0$		100	100

■ In the hospital		Pneumonia	
		Yes	No
Alcoholism	Yes	20	10
	No	80	90
OR = $\frac{20 \times 90}{80 \times 10} = 2.25$		100	100

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Las formas de evitar este sesgo en los estudios de cohortes es definir los criterios de selección de expuestos y no expuestos independientemente de los efectos (enfermedad, desenlace), y en los estudios de casos y controles, aplicando criterios de selección de enfermos (casos) y no enfermos (controles) independientemente de las exposiciones y, cuando es posible, extraer casos y controles de diferentes fuentes. Otra forma de evitarlos es realizando, cuando es posible, ensayos controlados aleatorizados (Capítulo 3) u otros estudios observacionales, como se verá en el Capítulo 7.

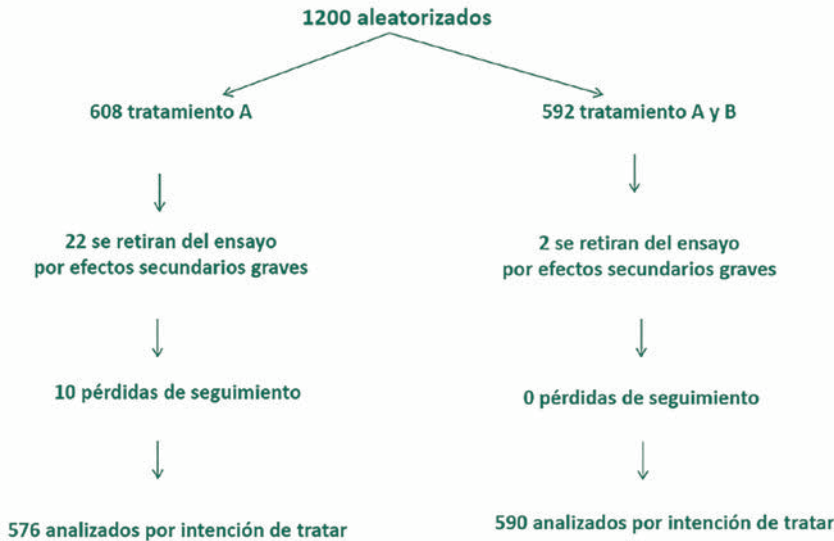
El **sesgo de información** se produce cuando también se distorsiona la magnitud de la asociación entre dos variables (como en el caso anterior), pero aquí como consecuencia de haber recabado la información de distinta forma (o que sea mejor, más fiable, más completa) para los expuestos y no expuestos o para casos y controles.

Los **sesgos de clasificación errónea** (*misclassification bias*) mencionados antes son un claro exponente de este tipo de sesgo y se dividen en dos tipos:

- Diferencial: se produce cuando el grado de clasificación errónea es mayor en un grupo que en otro (por ej., en casos que en controles o viceversa) y
- No diferencial: se introduce cuando el grado de clasificación errónea no difiere entre ellos.

Otro ejemplo de sesgo de información es el que se produce cuando los resultados de un estudio de cohortes o de un ensayo clínico se interpretan omitiendo que se hayan perdido en el seguimiento más participantes en uno que en otro grupo o en ambos de forma notable. Como se indicó en el Capítulo 3 con los ensayos clínicos, en esto estriba la importancia de incluir en los estudios los diagramas de flujo de los participantes (ilustrado en el artículo de Harrison CN, et al. en dicho Capítulo: [Harrison CN, et al. Ruxolitinib versus Best Available Therapy for ET intolerant or resistant to Hydroxycarbamide in a Randomized trial. Blood. 2017;26;30:1889-97.](#)), porque indican sobre los participantes que iniciaron el

estudio en cada grupo cuántos lo acabaron, cuántos se retiraron (withdrawals), cuántos se perdieron en el seguimiento (lost to follow up) y, lo más importante, los motivos. Si el desequilibrio entre grupos es marcado, los resultados pueden estar sesgados:



El sesgo de selección puede introducirse asimismo en revisiones de la bibliografía y en metanálisis y en este hecho estriba la importancia de que en estos estudios también se incluyan diagramas de flujo. Los resultados finales pueden ser distintos en función de los estudios que se incluyan u omitan. Por ejemplo, un sesgo que se baraja en estos estudios es el de la Torre de Babel (*Tower of Babel bias*), que se produce cuando sólo se incluyen estudios publicados en un idioma (generalmente el inglés o el castellano) y se omiten otros, escritos en otros idiomas y de diferente calidad, cuya inclusión se comprueba que cambia las conclusiones. Al realizar o reevaluar revisiones sistemáticas es muy útil seguir la guía PRISMA, que puede bajarse de internet: www.equator-network.org.

Y, por último, otro sesgo es el **de recuerdo** (*recall bias*): los casos acostumbran recordar mejor que los controles hechos pasados importantes (sobre la exposición, relevantes para el estudio) y los resultados basados en esta información diferencial pueden estar sesgados. Si hay controles que recuerdan menos exposiciones pasadas que los casos (éstos, por su enfermedad tienden a recordar más eventos pasados). Cuando nos encontremos en estas situaciones siempre hemos de hacer el siguiente ejercicio para apreciar los efectos de sesgo en los resultados (en los efectos estimados): averiguar cómo se modifican los valores de la tabla 2 x 2 correspondiente, como se ha razonado anteriormente. Por ejemplo, en el caso que acabamos de comentar, en la tabla 2 x 2 parte de los controles expuestos (b) pasarán, como consecuencia del sesgo, a ser no expuestos (d) (no lo recuerdan y se consideran no expuestos) y, por ello, el OR ($a \times d / b \times c$) aumenta, porque (d) aumenta en el numerador y b disminuye en el denominador, es decir, queda sobreestimado respecto al OR real. Este sesgo también viene determinado por el tiempo que haya transcurrido entre la exposición y la realización del estudio y de la gravedad de la enfermedad.

Las siguientes son estrategias para minimizar la introducción de un sesgo de información:

- Utilizar rigurosamente un protocolo estandarizado de obtención de datos para todos los participantes, que impida obtener información diferente para expuestos y no expuestos y para casos y controles.
- Garantizar que las fuentes y los métodos para recabar la información no difieren entre ambos grupos.
- Asegurarse de que los investigadores (incluidos los entrevistadores) desconocen o no se dejan influir por el conocimiento de la exposición y del efecto.
- Adoptar siempre medidas para valorar su existencia durante el diseño, la realización, el análisis y la interpretación de los resultados del estudio, y siempre comunicarlo cuando se detecte.

6. Otras aplicaciones de los estudios observacionales

Acabemos señalando que, como se indicó en el Capítulo 3, la evaluación y la vigilancia postautorización de la efectividad y la seguridad de medicamentos nuevos autorizados (o dispositivos y aparatos) están cobrando protagonismo, afortunadamente, cada vez con más intensidad. Esto explica que en los últimos años esté aumentando la utilización de **estudios observacionales postautorización** de medicamentos, de dispositivos y aparatos médicos, de pruebas diagnósticas, intervenciones quirúrgicas y de otras intervenciones para estimar sus efectos utilizando **datos de la vida real** (*Real World Data*). Los estudios observacionales de **efectividad comparada** (*comparative effectiveness*) con datos de la vida real y la farmacovigilancia persiguen, junto con la aplicación de nuevos métodos, aunque de forma sucinta, este fin: recopilar información adicional sobre efectividad y seguridad en condiciones habituales de práctica clínica (no experimentales) y compensar así el **margen de incertidumbre siempre asociado con los resultados de los ensayos clínicos**, la brecha entre eficacia y efectividad, siempre y cuando, obviamente, el diseño, su realización, los análisis, los resultados y su interpretación sean correctos conforme al paradigma actual de la investigación.

Como se ha comentado, con esos otros diseños de estudios observacionales, si se cumplen ciertos requisitos específicos de cada uno de ellos, se pueden **construir** grupos de estudio o intervención y grupos control similares a los que se obtienen por medio de la asignación aleatoria. Hace muchos años que se emplean en economía y su uso en medicina está aumentando lentamente, sobre todo a raíz de la importancia que está cobrando la efectividad comparada. Como es crucial, cuando menos, tener un conocimiento muy básico de ellos, se describirán brevemente en el capítulo 7.

7. Bibliografía

General

Nivel introductorio

- Gordis L. Epidemiology, Second Edition. Philadelphia, PA: WB Saunders Company; 2000
- Mausner JS, Kramer S. Epidemiology: An introductory text. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. Foundations of epidemiology. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epidemiología clínica. Ciencia básica para la medicina. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. Epidemiology, Beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc. 2000.

Específica de este Capítulo

En las referencias del nivel introductorio y del intermedio hay capítulos que tratan estos temas en detalle y con ejemplos.

- Breslow NE, Day NE. Statistical methods in cancer research. Vol I: The analysis of Case control studies. Lyon: International agency for research on cancer; 1980.
- Breslow NE, Day NE. Statistical methods in cancer research. Vol II: The design and analysis of cohort studies: Lyon: International agency for research on cancer; 1987.
- Schlesselman JJ. Case-control studies. Design, Conduct, Analysis. New York: Oxford University Press; 1982.

Capítulo 5.

Confusión e interacción

Índice

1. Confusión.....	74
1.1. Definición y ejemplos	74
1.2. ¿Qué se debe hacer con las variables de confusión?	77
2. Interacción	79
2.1. Definición y ejemplos	79
2.2. ¿Cómo actuar cuando se detecta interacción?	81
3. Confusión e interacción en los modelos multivariantesción	82
4. Bibliografía.....	84

Objetivos específicos de este Capítulo

1. Aprender el significado de los conceptos confusión e interacción.
2. Conocer sus implicaciones en los diferentes tipos de estudios y saber cómo actuar cuando se detectan.
3. Entender desde el punto de vista conceptual cómo se tratan la confusión y la interacción en los modelos multivariantes.

Nota importante

En este Cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I, 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como **Material didáctico complementario**.

1. Confusión

1.1. Definición y ejemplos

Recordemos que cuando se encuentra una asociación entre variables siempre tenemos que intentar averiguar si es cierta o si es, por el contrario, espuria, falsa, como consecuencia del azar (error aleatorio, por ejemplo, por efecto del muestreo) o de errores sistemáticos (sesgos (*biases*) o de factores de confusión (*confounding*)).

En términos generales, el valor estimado de un parámetro (por ej., en este caso, el RR o el OR) puede definirse del siguiente modo:

$$\text{Valor estimado} = \text{valor verdadero} + \text{error aleatorio} + \text{error sistemático}$$

Sólo en ausencia de errores aleatorios y sistemáticos (sesgos o factores de confusión) el valor estimado se corresponde con el valor real. **Descartar sesgos y confusión es la actividad más importante, y difícil, al investigar causalidad.** Al haber visto en el Capítulo 4 los sesgos, centrémonos ahora en la confusión.

La confusión, un concepto esencial en el estudio de la causalidad, se produce cuando una asociación encontrada entre exposición y resultado (desenlace, efecto, enfermedad, *outcome*) no se corresponde con la verdad, porque es artefactual a causa de la existencia de una tercera **variable** (o de más de una) que se denomina **confusora**.

Para que una variable C (por ejemplo, la edad) sea confusora de una asociación entre una exposición y una enfermedad (efecto o desenlace), C tiene que:

- (1) estar asociada con el efecto y, simultáneamente,
- (2) estar asociada con la exposición.

Veamos un ejemplo. En un estudio de 100 casos y 100 controles, 30% de los casos y 18% de los controles estuvieron expuestos al supuesto factor de riesgo o supuesta causa de la enfermedad. La fuerza de la asociación OR = 1,95.

Exposed	Cases	Controls
Yes	30	18
No	70	82
Total	100	100

$$OR = \frac{30 \times 82}{70 \times 18} = 1.95 \quad \text{Chi sq} = 3.95$$

Si con nuestros conocimientos sospechamos que la edad puede ser una variable confusora, comprobémoslo siguiendo los pasos señalados:

1. La edad tiene que estar asociada con el efecto:

Distribution of Cases and Controls by Age		
Age	Cases	Controls
< 40 years	50	80
≥ 40 years	50	20
Total	100	100

Los casos son mayores que los controles y, por consiguiente, se cumple el criterio 1.

2. La edad tiene que estar asociada con la exposición. Como se observa en la siguiente tabla, hay más expuestos entre los mayores de 40 años (50% frente 10%), lo cual demuestra que se cumple el criterio 2.

Relationship of Exposure to Age				
Age	Totals	Exposed	Not Exposed	Percent Exposed
< 40 years	130	13	117	10%
≥ 40 years	70	35	35	50%

Chi sq = 39.9

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Veámoslo con un ejemplo clásico en epidemiología:



CANAS (exposición) → MORTALIDAD (resultado)

	Mueren	No mueren
Canas	a	b
Sin canas	c	d

OR = 4,5

Los odds de morir en los que tienen canas son 4,5 más altos que en los que no las tienen y esta fuerza de asociación no es baja. Pero, como se ha señalado, descartemos que se trate de una asociación espuria y, tras descartar sesgos, pensemos si algún posible factor de confusión puede explicarla.

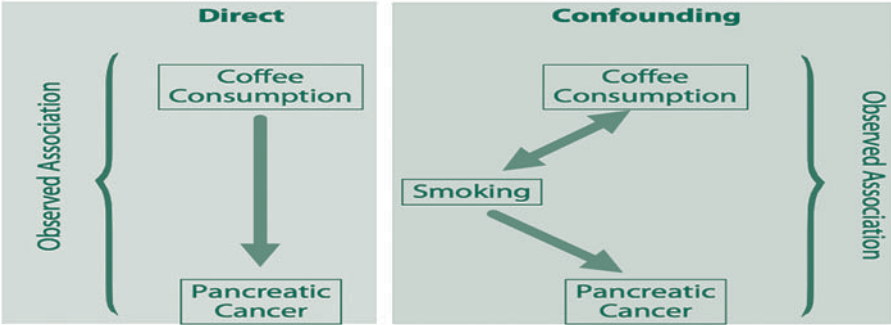
La edad es una clara variable de confusión (confusora, *confounder*, en econometría también se conoce como *omitted variable*, variable omitida), porque está asociada **simultáneamente** con las canas y la mortalidad. (Lo comprobaríamos al estratificar la tabla 2 x 2 como se ha hecho en el ejemplo anterior.)

Vinculando estos conceptos con lo estudiado en estadística, lo que muestra esta tabla 2 x 2 lo obtendríamos con un modelo logístico, cuya formulación en la ecuación que sigue no se corresponde con él sino con el de la regresión lineal, solamente por simplificar la exposición (la OR = e^{beta}):



Como se ilustra en este otro caso sobre la asociación entre el consumo de café y el cáncer de páncreas, el tabaco actúa como variable de confusión de esta asociación, porque está asociado *simultáneamente* con la exposición (es más frecuente que los que fuman consuman café que los que no fuman) y con el resultado (cáncer de páncreas).

Por este motivo, la asociación considerada supuestamente causal o de riesgo al inicio entre consumo de café y cáncer de páncreas es espuria, a resultas del efecto de una tercera variable de confusión, el consumo de tabaco, porque fumar está asociado tanto con el consumo de café como con el cáncer de páncreas.



1.2. ¿Qué se debe hacer con las variables de confusión?

Si la edad es una variable de confusión de una asociación, al estratificar dicha asociación por grupos de edad, la asociación desaparece. Si la asociación no desaparece, la edad no es una variable de confusión. Ante la duda, siempre debe seguirse este procedimiento (siempre hay que estratificar) con todas las variables que sospechemos que puedan ser factores de confusión y comprobar si hay o no confusión. Si tuviéramos los intervalos de confianza de los dos OR de la siguiente tabla, en este caso habría solapamiento entre sus valores.

Calculations of Odds Ratios in a Stratified Analysis				
Age	Exposed	Cases	Controls	Odds Ratios
< 40 years	Yes	5	8	$\frac{5 \times 72}{45 \times 8} = \frac{360}{360} = 1.0$
	No	45	72	
	Total	50	80	
≥ 40 years	Yes	25	10	$\frac{25 \times 10}{25 \times 10} = \frac{250}{250} = 1.0$
	No	25	10	
	Total	50	20	

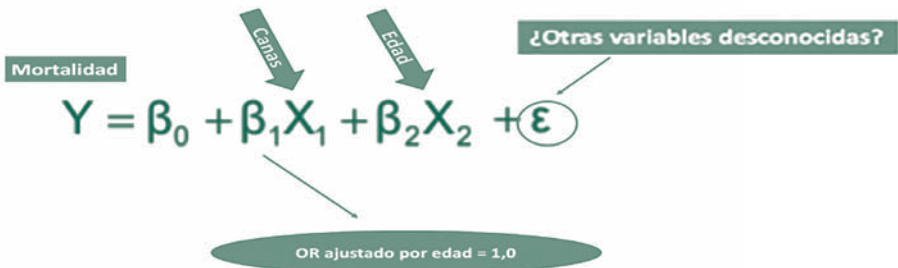
<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Estratificando por edad en el caso de la asociación entre canas y mortalidad comprobamos que la asociación desaparece y, por ello, la edad es una variable de confusión.

		Mueren	No mueren	OR = 1,0
Jóvenes	Canas	a	b	
	Sin canas	c	d	
		Mueren	No mueren	OR = 1,0
Mayores	Canas	a	b	
	Sin canas	c	d	

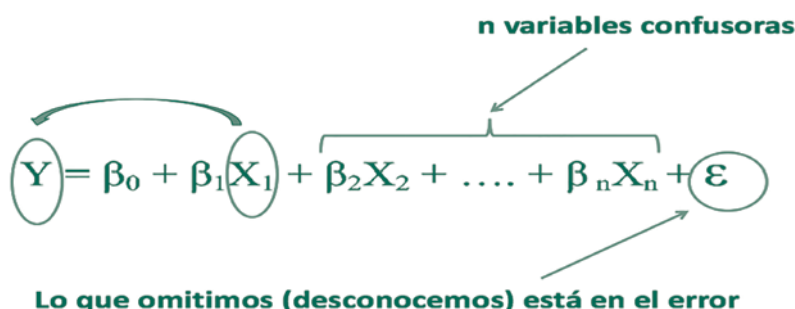
El OR global de la tabla también sería 1,0. Este OR global para los dos estratos de edad, de la forma como se calcula, es el OR de la asociación entre canas y mortalidad ajustado por edad.

En la regresión, para hacer lo mismo que con la estratificación, añadiríamos la edad como variable independiente (x_2). Entonces, se observa que el OR de la asociación entre canas y mortalidad también desciende a 1,0. Este OR está ajustado por edad. En la regresión se consigue el mismo ajuste que con la estratificación:



Los modelos multivariantes permiten ajustar por varias variables simultáneamente, lo cual sería harto engorroso hacerlo “a mano” estratificando con tablas 2 x 2, como se ha hecho en la tabla anterior. En un modelo multivariante, la fuerza de la asociación (RR, OR o HR, que se derivan de los coeficientes de regresión beta) entre cada variable independiente y la dependiente (por ejemplo, en el modelo de abajo entre X_1 e Y) está ajustada por todas las restantes variables independientes incluidas en el modelo y así se elimina la capacidad de todas estas variables de *confundir* las asociaciones detectadas entre cada una de ellas (como variables independientes) y la dependiente del modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

n variables confusoras

Lo que omitimos (desconocemos) está en el error

En estos ejemplos trabajamos con regresiones logísticas y en ellas el número e^{beta} (el coeficiente de regresión) es el OR, pero la interpretación de la beta es distinta, como se ha visto en esta-
dística.

Fijémonos en el **término de error, esencial en todos los modelos y al que se presta poca atención**: expresa lo que “aún le queda” al modelo, con las variables independientes que se hayan incluido, para poder predecir o explicar al 100% la variable dependiente Y , es decir, lo que desconocemos o no hemos medido, las variables no observadas (*unobservables*) que comentábamos en el Capítulo 3, nuestras omisiones, nuestro desconocimiento (ignorancia) o ambas cosas, que se conoce como **confusión residual** (*residual confounding*). Esto ocurre en muchísimos estudios: es muy difícil, por no decir virtualmente imposible, descartarla completamente. (En el capítulo 6 Parte II se ofrece más información sobre los errores de los modelos.)

La confusión debida a *observables* se elimina ajustando los estimadores de la fuerza de las asociaciones con análisis estratificados, con modelos multivariantes como hemos visto o emparejando los integrantes del grupo de intervención con los del grupo control en función de estas variables de confusión o por medio de la **asignación aleatoria**, el método más afectivo para hacerlo, porque distribuye todas las variables, observadas y no observadas (conocidas y desconocidas, confusoras o no) de forma muy parecida o igual entre los grupos experimental y control. **En ausencia de asignación aleatoria, lo que ocurre con todos los estudios observacionales, buscar y eliminar sesgos y la confusión con diseños y métodos adecuados debe ser la tarea principal para poder estimar efectos causales reales exentos de la influencia de ellos.** En el Capítulo 7 ampliaremos las explicaciones de estos conceptos.

De ahí el gran esfuerzo que debemos hacer siempre en pensar en posible confusores y la humildad y la cautela con que hay que proceder al estimar efectos causales, extraer conclusiones y al interpretar las que otros extraen.

2. Interacción

2.1. Definición y ejemplos

Hemos dicho que, cuando encontramos una asociación entre variables, debemos preguntarnos si es real o si, por el contrario, es espuria a consecuencia del azar (errores aleatorios), sesgos o factores de confusión (errores sistemáticos).

Si la asociación parece existir, tras haber descartado justificadamente sesgos y confusión y, además, la estratificamos en función de una tercera variable, hemos de preguntarnos: ¿varía la fuerza de la asociación entre los estratos, estimada con RR o HR en estudios de cohortes o en ensayos clínicos o con OR en estudios de casos y controles?, es decir, ¿varía la fuerza de la asociación al variar los niveles o valores de la tercera variable? Si lo hace, hay interacción; si no varía, se descarta la interacción.

En la siguiente tabla se presentan las incidencias de cáncer de pulmón de mineros del uranio expuestos y no expuestos al tabaco y al radón. Se observa que la incidencia es mucho mayor en los expuestos a ambos simultáneamente.

Smoking	Radon	Lung Cancer Incidence
No	No	1/1000
No	Yes	5/1000
Yes	No	10/1000
Yes	Yes	50/1000

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Si no hay interacción entre la exposición al radón (R) y al tabaco (T), la incidencia esperada (Iê) de cáncer en los expuestos (exp) a ambos sería 14:

Incidence		Smoking	
		-	+
Radon	-	1	10
	+	5	50

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

$$I_{\hat{e} \text{ exp a R y T}} = (I_{\text{exp a R}} - I_{\text{no exp a R}}) + (I_{\text{exp a T}} - I_{\text{no exp a T}}) + I_{\text{en no exp a R ni a T}} =$$

$$(5-1) + (10-1) + 1 = 14$$

pero en los expuestos al radón y al tabaco es mucho más alta, 50, y, por ello, hay interacción, que en este caso es multiplicativa, más que aditiva (la incidencia observada en los expuestos a ambos es 50, mucho mayor que la suma de ambas, 14).

Cuando el riesgo estimado (incidencia, RR) de desarrollar una enfermedad al estar expuesto a los dos factores de riesgo A y B supuestos es igual al esperado, no hay interacción, como en el siguiente ejemplo:

Interacción con incidencias

1. Incidence		Factor A	
		-	+
Factor B	-	3	9
	+	15	21

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

El riesgo esperado en este caso utilizando incidencias es:

$$(15-3) + (9-3) + 3 = 21 = \text{al observado}$$

Y cuando el riesgo de los que están expuestos a los dos factores de riesgo es menor que el esperado, hay interacción antagónica, como en este ejemplo:

Incidence		Factor A	
		-	+
Factor B	-	1.0	4.0
	+	3.0	2.0

El riesgo esperado en este caso es:

$$(3 - 1) + (4 - 1) + 1 = 6$$

es decir > 2, el observado.

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

Interacción con RR u OR

Para comprobar si hay o no interacción, cuando trabajamos con RR u OR en lugar de con incidencias, se procede del siguiente modo:

Relative Risk		Factor A	
		-	+
Factor B	-	1.0	3.0
	+	5.0	15.0

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

En este caso, como el RR esperado en los expuestos a A y B es $5,0 \times 3,0 = 15,0$, coincide con el observado, no hay interacción.

Pero en el siguiente caso, el RR esperado en los expuestos a A y B = $1,23 \times 1,53 = 1,88$: Por consiguiente, como el observado $5,71 > 1,88$, concluimos que hay interacción.

Relative Risk		Smoking	
		No	Yes
Alcohol	No	1.00	1.53
	Yes	1.23	5.71

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/fundepiii/coursePage/lectureNotes/>

2.2. ¿Cómo actuar cuando se detecta interacción?

Es sumamente importante que quede claro lo siguiente: **si hay interacción, los resultados han de presentarse estratificados**, no debe presentarse solo la estimación global de la fuerza de asociación global, porque sólo la global oculta la interacción: el hecho de que la incidencia o la fuerza de la asociación varía según el valor o nivel de una tercera variable.

En el siguiente caso, la estratificación de la asociación entre la enfermedad D y la exposición E revela que, como los RR global y los RR de cada estrato son iguales, la variable C por la cual se estratifica no es confusora ni existe interacción.

Stratum 1 (C+)				Stratum 2 (C-)				Pooled			
		D+	D-			D+	D-			D+	D-
E+	160	240	400	E+	40	560	600	E+	200	800	1000
E-	40	360	400	E-	10	590	600	E-	50	950	1000

$RR_1 = (160 / 400) / (40 / 400) = 4.0$
 $RR_2 = (40 / 600) / (10 / 600) = 4.0$
 $cRR = (200 / 1000) / (50 / 1000) = 4.0$

<http://www.sjsu.edu/faculty/gerstman/StatPrimer/stratified>

En este otro caso se observa, por el contrario, que el RR global = 4,0, al estratificar la asociación entre E y D por una tercera variable C, que sospechamos es confusora, la asociación desaparece, puesto que los RR de ambos estratos = 1,0, lo cual confirma que C es una variable de confusión (y, obviamente, no hay interacción).

Stratum 1 (C+)				Stratum 2 (C-)				Pooled			
		D+	D-			D+	D-			D+	D-
E+	194	606	800	E+	6	194	200	E+	200	800	1000
E-	24	76	100	E-	26	874	900	E-	50	950	1000

$RR_1 = (194/800) / (24/100) = 1.0$
 $RR_2 = (6 / 200) / (26 / 900) = 1.0$
 $cRR = (200/1000) / (50 / 1000) = 4.0$

<http://www.sjsu.edu/faculty/gerstman/StatPrimer/stratified>

Por último, en el siguiente caso, al estratificar la asociación entre E y D, cuyo RR global = 4,0, vemos que los RR (la fuerza de dicha asociación) varía en función de los valores de una tercera variable C y concluimos, por definición, que hay interacción.

Stratum 1 (C+)				Stratum 2 (C-)				Pooled			
	D+	D-			D+	D-			D+	D-	
E+	12	188	200	E+	188	612	800	E+	200	800	1000
E-	48	752	800	E-	2	198	200	E-	50	950	1000

$RR_1 = (12 / 200) / (48/800) = 1.0$
 $RR_2 = (188/800) / (2/ 200) = 23.5$
 $cRR = (200/1000) / (50 / 1000) = 4.0$

<http://www.sjsu.edu/faculty/gerstman/StatPrimer/stratified>

Como en la etiología de muchas enfermedades suelen desempeñar algún papel causal varios factores de riesgo (actuando simultáneamente, como el paso del tiempo va demostrando a medida que avanza la investigación en diversas áreas de las ciencias de la salud), es muy importante entender el concepto de interacción estadística, pero a sabiendas de que su explicación no es estadística, sino que pertenece a la esfera de la biología y la fisiopatología, la economía o las ciencias sociales, según el problema que se analice.

3. Confusión e interacción en los modelos multivariantes

En los modelos multivariantes también puede comprobarse si existe o no interacción, introduciendo en el modelo el **término de interacción** entre las variables implicadas. Sean en el siguiente modelo x_1 la edad y x_2 el sexo:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \epsilon$$

Y es la variable dependiente, β_0 es la constante del modelo, β_1 es el coeficiente de regresión, en este caso hipotético, de la edad, β_2 , el del sexo, y β_3 , el del término de interacción entre la edad y el sexo (X_1X_2). Aquí, concretamente, nos preguntamos si la fuerza de la asociación entre la edad y la variable dependiente varía en función del sexo. Si el coeficiente β_3 es estadísticamente significativo, indica que hay interacción, es decir, que la asociación entre la edad e Y varía en función del sexo, no es igual en hombres que en mujeres, y si β_3 no es significativo, que no existe interacción, es decir, que la fuerza de la asociación entre la edad e Y es la misma en los dos sexos. En estos modelos también pueden introducirse términos de interacción entre más de dos variables, pero su interpretación se torna mucho más compleja.

En cualquier caso, y se debe insistir en ello: **si hay interacción, los resultados han de presentarse estratificados, no debe presentarse solo la estimación global de la fuerza de asociación global, porque la global oculta la interacción y dicha interacción tiene un significado importante.**

Si la edad (X_1) fuese una variable de confusión de la asociación entre el sexo (X_2) y la variable dependiente (Y), en estos modelos, como incluyen como variables independientes tanto el sexo como la edad, la asociación entre el sexo e Y está ajustada por edad, lo cual elimina el efecto confusor de la edad. Si el modelo tuviese más variables independientes, todas las asociaciones entre cada variable independiente y la Y estarían ajustadas por todas esas variables.

La inclusión en los modelos de términos de interacción significativos reduce la magnitud del término de error y aumenta, por ende, la capacidad predictiva, explicativa o discriminante de Y del modelo.

Para finalizar, pongamos un ejemplo ilustrativo real. Cada vez se estima más la efectividad y seguridad de nuevas intervenciones quirúrgicas en estudios observacionales en los cuales se comparan con las de las usadas hasta la fecha. En ellos se construyen diferentes modelos para estimar efectos de cada técnica quirúrgica (variable independiente) y cada variable de resultado Y (por ejemplo, en cirugía oncológica, para cada técnica quirúrgica se puede estimar mortalidad, para período libre de enfermedad, supervivencia global, recidivas, presencia de tumor en bordes resecaos, ganglios extraídos e invadidos, complicaciones intra y postoperatorias, fallo en el rescate, etc., con un modelo para cada Y).

Como variables independientes se incluye cada tipo de cirugía estudiado (en una variable *dummy*). Su beta mide la asociación entre cada tipo de cirugía y la variable Y de que se trate en cada modelo. También han de incluirse en el modelo numerosas variables potenciales de confusión de esta asociación (se comprueba si lo son o no y las que se confirman se quedan en el modelo), como el estadio funcional de los pacientes, edad, sexo, IMC, comorbilidad, tipo de tumor, volumen, estadio, grado de diferenciación celular, el cirujano, muchas veces omitido por error (también como variable *dummy*) y un posible largo etcétera.

Un término casi nunca incluido, cuya omisión puede constituir un error importante, es el de interacción entre el número de intervenciones realizadas por cada cirujano y el tiempo (en meses o años) durante el cual cada cirujano haya hecho las intervenciones incluidas en el estudio. Este término de interacción captura la curva de aprendizaje de los cirujanos: es lógico esperar que si en un estudio se incluyen, digamos, 100 intervenciones por cirujano, éste tenga más pericia y obtenga mejores resultados en la 80 y la 100 que en la 10 y 20, siempre ajustando por las restantes variables incluidas en el modelo. La inclusión de este término, si la interacción es significativa, contribuye a reducir el término de error, a ajustar mejor las restantes asociaciones del modelo, a mejorar su capacidad predictiva o explicativa global de la Y , así como a mejorar las estimaciones de los efectos causales del tipo de cirugía, las diferentes Y enumeradas antes a modo de ejemplo.

4. Bibliografía

General

Nivel introductorio

- Gordis L. Epidemiology, Second Edition. Philadelphia, PA: WB Saunders Company; 2000.
- Mausner JS, Kramer S. Epidemiology: An introductory text. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. Foundations of epidemiology. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epiemiología clínica. Ciencia básica para la medicina. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. Epidemiology, Beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc.; 2000.

Específica de este Capítulo

- En las referencias del nivel introductorio y del intermedio hay capítulos que tratan estos temas en detalle y con ejemplos.

Capítulo 6. Pruebas diagnósticas. Parte I

Índice

1. Introducción	86
2. Criterios de validación de las pruebas diagnósticas	86
3. Validez clínica	87
4. Utilidad clínica	93
5. Algunas cautelas	94
6. Material didáctico complementario	95
7. Bibliografía	95

Objetivos específicos de este Capítulo

1. Conocer los criterios de validación de las pruebas diagnósticas (validez técnica, validez analítica y utilidad clínica).
2. Aprender a aplicarlos e interpretarlos en la práctica clínica y conocer los errores por omisión y comisión más frecuentes en su aplicación clínica habitual.
3. Conocer la utilidad de dichos criterios para la regulación (autorización) de pruebas diagnósticas, su financiación e incorporación en la práctica clínica.
4. Aprender a revisar sistemáticamente los estudios de validación de pruebas diagnósticas y conocer las principales deficiencias de los publicados.

Nota: en este módulo no se explicará el cribado (*screening*) de enfermedades.

Nota importante

En este Cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I, 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como **Material didáctico complementario**.

1. Introducción

La evaluación rigurosa de las pruebas diagnósticas debería ser crucial para que:

1. el regulador decida si deben autorizarse o no,
2. el financiador decida su nivel de cobertura y negocie su precio con la industria teniendo en cuenta el *valor diagnóstico añadido* que aportan las nuevas,
3. el gestor de servicios decida si debe priorizarlas e incorporarlas o no en los centros asistenciales, y
4. el clínico decida si tiene que incorporarlas o no a su arsenal diagnóstico en la práctica clínica y, en caso afirmativo, cómo utilizarlas.

Por otro lado, la evaluación final del *valor* de una prueba diagnóstica debe basarse en:

1. la medida en que el estudio de su validación incluye toda la información necesaria para evaluar su validez interna y externa,
2. que dicha información sea apropiada conforme a los estándares metodológicos,
3. que los resultados que aporta sean metodológicamente correctos,
4. que cumplan con los principales estándares regulatorios, es decir su *validez técnica*, su *validez analítica* y su *utilidad clínica*, y
5. que tenga *valor diagnóstico incremental* (añadido, relativo) respecto a las disponibles.

No obstante, numerosas evaluaciones publicadas ponen de manifiesto, en primer lugar, que muchos estudios realizados que intentan demostrar la validez de las pruebas diagnósticas incumplen estándares de calidad específicos para estas pruebas, omiten información mínima esencial para poder evaluar su validez o esta información se circunscribe a uno de los tres parámetros clave de su evaluación (su validez clínica) y soslayan los dos restantes (su validez técnica y su utilidad clínica).

En segundo lugar, que la regulación de las pruebas diagnósticas ha sido deficiente durante décadas, porque los estándares científicos que se les han exigido durante años para autorizarlas no han sido estrictos, sino insuficientes y mucho más laxos que los de los medicamentos y los de los aparatos y dispositivos médicos. Ello ha desembocado recientemente en una revisión profunda de sus criterios de autorización y en la aprobación de nuevas directivas en la Unión Europea y en la revisión de los criterios de autorización por la FDA en los EEUU.

2. Criterios de validación de las pruebas diagnósticas

Los tres criterios clave y esenciales de la evaluación de una prueba diagnóstica son:

Validez técnica: es la capacidad de la prueba para detectar o medir lo que pretende medir o detectar (una estructura con determinado tamaño y otras características por

imagen, una mutación, una secuencia de genes, la concentración de un metabolito en algún fluido o tejido, la captación por ciertas células de un metabolito marcado visible por imagen...).

Validez clínica: expresa el grado en que la prueba detecta correctamente a las personas con determinada enfermedad o característica y a las que no la tienen o si un biomarcador de respuesta discrimina correctamente qué pacientes responderán a un medicamento y cuáles no responderán o si un biomarcador de toxicidad permite discernir en cuáles será tóxico el medicamento y en cuáles no, y la precisión, exactitud y fiabilidad con que lo hacen.

Utilidad clínica: indica si, con su validez técnica y su validez clínica, incide o no en la mejora de los resultados en salud finales de la atención clínica que recibe el paciente y si los beneficios que aporta superan a sus riesgos, en definitiva, si contribuyen en alguna medida a mejorar dichos resultados.

En este Capítulo se explicarán la validez clínica y la utilidad clínica, no la validez técnica, propia de cada prueba, porque su carácter técnico e instrumental no es un objetivo de la epidemiología.

3. Validez clínica

Los elementos clave e imprescindibles consensuados a escala internacional que deben utilizarse para evaluar la validez interna y externa de y elaborar, respectivamente, un estudio de validación de pruebas diagnósticas se encuentran en dos guías, QUADAS y STARD. A estas guías y a las de otros tipos de estudios se accede libremente aquí: <http://www.equator-network.org/>.

Si se soslaya alguno de los 5 criterios para evaluar el valor de una prueba diagnóstica enumerados más arriba, la evaluación será incompleta y las decisiones basadas en ella pueden desembocar en errores clínicos, sub o sobrediagnóstico, posibles eventos adversos y menor eficiencia. Recordémoslos:

1. la medida en que el estudio de su validación incluye toda la información necesaria para evaluar su validez interna y externa,
2. que dicha información sea apropiada conforme a los estándares metodológicos,
3. que los resultados que aporta sean metodológicamente correctos,
4. que cumplan con los principales estándares regulatorios, es decir su *validez técnica*, su *validez analítica* y su *utilidad clínica*, y
5. que tenga *valor diagnóstico incremental* (añadido, relativo) respecto a las disponibles.

Seguidamente se definen los **parámetros en que se basa la valoración de la validez clínica de una prueba diagnóstica**. Pasaremos revista a ellos con rapidez, porque son

conceptos muy conocidos y utilizados por todos, aunque haremos hincapié en ciertos aspectos que con frecuencia se omiten o interpretan erróneamente.

En la siguiente tabla 2 x 2 se definen los verdaderos positivos y negativos y los falsos positivos y negativos de una prueba (*test outcomes*) respecto al estándar de oro (*gold standard*, es decir la mejor prueba diagnóstica disponible). Asimismo, se definen los cuatro parámetros de validez clínica más empleados.

		Presencia de la enfermedad o atributo confirmado con el <u>gold standad</u>		
		Positivo	Negativo	
Positivo		verdadero positivo (A)	falso positivo (B)	Valor predictivo positivo = $A / A + B$
		falso negativo (C)	verdadero negativo (D)	
Negativo				Valor predictivo negativo = $D + C + D$
		Sensibilidad = $A / A + C$		Especificidad = $D / B + D$

Sensibilidad (S, *sensitivity*) = capacidad de la prueba para detectar personas con la enfermedad o característica de interés o proporción de verdaderos positivos (*true positive rate*) o probabilidad de que la prueba sea positiva en los que la presentan.

Especificidad (E, *specificity*) = capacidad de la prueba para detectar a las personas sin la enfermedad o característica de interés o proporción de verdaderos negativos (*true negative rate*) o probabilidad de que sea negativa en los que no la presentan.

La sensibilidad y la especificidad nos ayudan en una primera consulta a decidir qué pruebas pedimos para diagnosticar una enfermedad tras haber recabado información sobre el paciente *a priori*.

Valor predictivo positivo (VPP, *positive predictive value*) = proporción de los resultados positivos de la prueba que corresponden a personas con la enfermedad o con la característica de interés o probabilidad de que los resultados positivos correspondan a personas que padecen la enfermedad o tienen dicha característica.

Valor predictivo negativo (VPN, *negative predictive value*) = proporción de los resultados negativos que corresponden a personas sin la enfermedad o la característica de interés o probabilidad de que los resultados negativos correspondan a personas que no padecen la enfermedad o no presentan dicha característica.

Los valores predictivos los usamos en la consulta sucesiva para interpretar los resultados de las pruebas solicitadas en la primera consulta. Cuando aumenta la prevalencia, el VPP aumenta y el VPN desciende.

En rigor, las estimaciones de estos cuatro parámetros se han de acompañar de sus IC95% para dar idea del error muestral asociado con ellas, pues siempre se estiman en muestras.

Hay un hecho de gran importancia que se olvida con mucha frecuencia: los valores de estos cuatro parámetros son estimaciones que hacen referencia a grupos de individuos (se han estimado en ellos) y no permiten discriminar a nivel de cada individuo. Es decir, por ejemplo, con una S y una E del 85%, sabemos que la prueba dará positiva y negativa, respectivamente, en 85 de cada 100 personas con y sin la enfermedad. En ellas, asimismo, un 15% de los resultados positivos y negativos serán falsos. Sin embargo, cuando una persona entra en la consulta y le solicitamos la prueba, no sabremos en ella, concretamente, como individuo, si el resultado positivo de la prueba es verdadero o falso positivo o si el resultado negativo de la prueba es negativo verdadero o negativo falso. En otras palabras, **ni la sensibilidad ni la especificidad tiene capacidad de discernir o discriminar sin error la presencia o ausencia de una característica o enfermedad a nivel de individual.**

Para aproximarnos más la probabilidad con que la prueba diagnóstica es capaz de detectar a una persona concreta con o sin la enfermedad o característica de interés, es decir, a la probabilidad post-prueba en un individuo, hemos de conocer:

1. primero, la **probabilidad pre-prueba de la enfermedad** (una estimación que hacemos, incluso intuitivamente en ausencia de más información, basándonos en *el conjunto* de signos (incluidos resultados de otras pruebas), síntomas y la prevalencia en la zona donde trabajamos es la mejor aproximación disponible a dicha probabilidad; aunque no dispongamos de estimaciones de todo ello exactas ni precisas, sí tenemos cierta idea, aunque sea *grosso modo*, de su magnitud) y
2. segundo, las **razones de verosimilitud** (RV, *likelihood ratios*) de una prueba positiva y una negativa, que se definen en la siguiente tabla de cuatro maneras distintas, para facilitar su comprensión:

Razón de verosimilitud positiva	Razón de verosimilitud negativa
Sensibilidad / (1 - Especificidad) (A / [A+C]) / (B / [B+D])	(1 - Sensibilidad) / Especificidad (C / [A+C]) / (D / [B+D])
% verdaderos positivos / % falsos positivos	% falsos negativos / % verdaderos negativos
Cuánto más verosímil es tener una prueba positiva en una persona con la enfermedad que en una sin la enfermedad	Cuánto más verosímil es tener una prueba negativa en una persona sin la enfermedad que en una con la enfermedad

No por ser menos conocidas que los restantes parámetros, las RV son menos importantes. El fundamento intuitivo de las dos razones de verosimilitud es bien sencillo: para “poder confiar” en una prueba, ha de ser mucho más verosímil (frecuente, probable) que la prueba sea positiva en los enfermos que en los sanos, y mucho más verosímil que sea negativa en los sanos que en los enfermos.

Repasemos: la **razón de verosimilitud de una prueba positiva (RV+)** es el cociente entre sus verdaderos positivos y sus falsos positivos o, también, el valor que nos dice cuántas veces es más verosímil que la prueba dé positiva en una persona con la enfermedad que en una sin la enfermedad. Por ejemplo, una $RV+ = 15$ significa que el 15 veces más verosímil que la prueba dé positiva en las personas con la enfermedad que en las que no la tienen. Se considera, por consenso, que las $RV+ < 10$ no permiten confirmar la enfermedad con confianza razonable y que, por encima de 10, cuanto mayor sea, mayor es dicha confianza.

Por su parte, la **razón de verosimilitud de una prueba negativa (RV-)** es el cociente entre sus falsos negativos y sus verdaderos negativos o, también, el valor que nos dice cuántas veces es más verosímil que la prueba dé negativa en los enfermos (falso negativo) que en los sanos (verdaderos negativos). Una $RV- = 0,8$ significa que por cada 10 verdaderos negativos habrá 8 falsos negativos. Una $RV- = 0,2$ significa que por cada 10 verdaderos negativos habrá sólo 2 falsos negativos, un resultado mucho más deseable. Y una $RV- = 0,1$, que por cada 10 verdaderos negativos habrá sólo 1 falso negativo, casi el ideal.

Por consenso se acepta a escala internacional que con **una $LR+ > 10$ la prueba es adecuada, más cuanto mayor sea, para confiar razonablemente en un diagnóstico, y que con una $LR- < 0,1$ es razonable para descartarlo, más cuanto menor sea.**

Fijémonos en que estas dos razones aportan más información que los cuatro parámetros anteriores, pues incluyen en su definición los cuatro valores de la tabla 2 x 2 (a, b, c y d), a diferencia de la S, la E, el VPP y el VPN, que sólo incluyen dos de ellos en su cálculo.

El valor principal de las RV es que al multiplicarlas por la probabilidad pre-prueba (la prevalencia) dan la probabilidad post-prueba de que un paciente concreto, a nivel individual, tenga la enfermedad. Esto es,

Probabilidad post-prueba = probabilidad pre-prueba x razón de verosimilitud

Una forma muy sencilla de hacer este cálculo en cada caso concreto es utilizar un **nomograma**, como los que se reproducen en la Figura 2 de artículo [Grimes DA, Shulz KF. Refining clinical diagnosis with likelihood ratios. Lancet. 2005;365:1500-5](#), cuya consulta es necesaria para entenderlo.

Para apreciar cómo las RV hacen que aumente o disminuya según su magnitud y signo la probabilidad post-prueba (al convertir la pre-prueba en probabilidad post-prueba), veamos la siguiente tabla, adaptada del mismo artículo anterior:

Cambio en la probabilidad de enfermedad				
RV		RV		
Cuanto mayor es de 1,0 la RV, mayor probabilidad de enfermedad		Cuanto menor es de 1,0 la RV, menor probabilidad de enfermedad		
10	45	1,0		0
9	40	0,5	↗	-15
5	30	0,4	↗	-20
4	25	0,3	↗	-25
3	20	0,2	↗	-30
2	15	0,1	↗	-45

Las RV entre 0 y 1 reducen la probabilidad de enfermedad post-prueba y las RV > 1, la aumentan. En otras palabras, **cuanto mayor es la RV+, más aumenta la probabilidad post-prueba, y cuando menor es la RV-, menor es la probabilidad post-prueba de enfermedad.**

Si se presta atención a las RV de **muchas pruebas diagnósticas**, supuestos factores de riesgo y biomarcadores considerados aisladamente (sin combinarlos con otra pruebas), se apreciará que en numerosas ocasiones las RV + de factores individuales (aislados) son mucho menores de 10 y las negativas, mucho mayores de 0,1, es decir, **no tienen capacidad para discriminar** adecuadamente desde un punto de vista clínico (no tienen capacidad para confirmar ni descartar) los que padecen y no padecen una enfermedad o tienen o no una característica de interés, lo cual es sumamente importante en clínica. (Los valores de las RV + y las RV- han de estimarse en tablas 2 x 2 o buscarse en la bibliografía y utilizarlos cuando, tras haber evaluado la calidad del estudio, se puedan considerar estimaciones aceptables de las RV.)

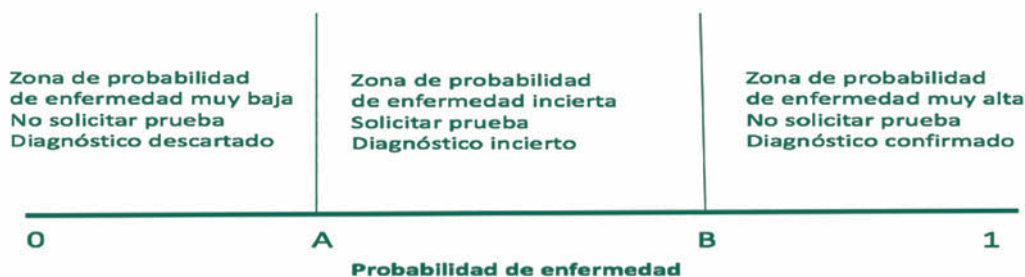
Un error por omisión frecuente es que en los estudios de validez de las pruebas diagnósticas no se estiman las RV, si bien desde hace pocos años está aumentando el número de estudios en que se incorporan estas estimaciones. Veamos un ejemplo que muestra la **baja capacidad discriminante de supuestos factores de riesgo aislados** (Campillo-Artero C, Serra-Burriel M, Calvo-Pérez A. Predictive modelling of emergency cesarean delivery. PLoS ONE. 13(1): e0191248. <https://doi.org/10.1371/journal.pone.091248>. Acceso libre).

Table 5. Prevalence ratios and positive likelihood ratios for the alleged risk factors for emergency c-sections

	Prevalence	95% CI	Prevalence eutocic	95% CI	Prevalence ECS	95% CI	Prevalence Ratio	SE	LR+	95% CI
Tobacco	12	12-12	12	12-12	13	12-14	1.104	0.0378	1.08	1-1.16
Previous C-sec.	11	11-11	10	10-10	22	20-24	1.028	0.0366	2.2	2-2.4
Comorbidity	17	17-17	17	16-18	25	23-27	1.585	0.0439	1.47	1.38-1.56
Obs. Risk	41	40-42	38	37-39	58	56-60	3.690	0.0568	1.52	1.47-1.57
Prev. Preg.	68	67-69	69	68-70	68	66-70	6.221	0.0524	0.98	0.95-1.01
Induction	23	22-24	20	19-21	43	41-45	2.062	0.0484	2.15	2.05-2.25
pH	9	9-9	7	7-7	29	27-31	2.636	0.0335	4.14	3.85-4.42
Night delivery	45	44-46	45	44-46	55	53-57	4.114	0.0576	1.22	1.17-1.26
BMI ≥25	41	40-42	37	35-39	51	49-53	3.611	0.0623	1.37	1.32-1.43
Age ≥35	27	26-28	26	25-27	34	32-36	2.441	0.0513	1.30	1.23-1.38

En este caso se presenta la estimada para la cesárea urgente (respecto al parto vaginal), que puede considerarse ejemplo muy representativo de lo que ocurre habitualmente con los supuestos factores de riesgo aislados. La RV+ más alta apenas es mayor de 4,0. Ningún supuesto factor de riesgo de los considerados en este ejemplo tiene capacidad para predecir o discriminar este tipo de parto en este ni en otros estudios; por este motivo no son factores de riesgo. **Es un error frecuente**, no exento de consecuencias si se actúa a partir de sus resultados, **considerar factor de riesgo a cualquier variable de exposición asociada con un resultado** (en este caso tipo de parto) **que, tras realizar estas estimaciones, muestra tener escasa capacidad discriminante.**

Ahora fijémonos en el siguiente esquema:



Fuente: adaptado de: Grimes DA, Schulz KF. *Lancet*. 2005;365:1500-5.

Las pruebas diagnósticas deberían solicitarse cuando van a modificar la actitud terapéutica (tratar o no). El valor de las pruebas diagnósticas es tanto mayor cuanto más nos encontramos entre A y B, la zona del diagnóstico incierto, zona con más incertidumbre. En esta zona se puede apreciar el valor de la magnitud de las RV para estimar cuánto son capaces de modificar la probabilidad pre-prueba (y aumentar la post-prueba por encima de B para confirmar el diagnóstico o reducirla por debajo de A, para descartarlo con un margen de confianza razonable). Esto es especialmente importante cuando la prueba que se duda pedir es invasiva y acarrea riesgo de daño para el paciente. Con un nomograma podemos averiguar cuán alta o baja ha de ser la RV para lograr estos cambios. Si con una prueba determinada no se puede conseguir traspasar A o B, no debe solicitarse, porque el valor diagnóstico que ofrece es bajo.

Conviene recordar que la mayoría de los estudios de utilización de pruebas diagnósticas realizados muestran que la *sobreindicación* es mucho más frecuente que la *subindicación*, y que este exceso de pruebas diagnósticas se ha asociado reiteradamente con **sobrediagnóstico** (*overdiagnosis*) y con los eventos adversos (evitables) asociados con las actitudes terapéuticas derivadas de él. Un ejemplo paradigmático de ello es el exceso de complicaciones evitables de las biopsias y las cirugías de próstata injustificadas por interpretar erróneamente los resultados de la PSA.

Tengamos en cuenta, además, que:

- el efecto de las RV en las probabilidades pre-prueba no es lineal: una RV = 100 no aumenta la probabilidad pre-prueba 10 veces más que una RV = 10,
- cuanto más cercanos a 1,0 son los valores de las RV, menos útiles son. Su utilidad aumenta cuanto más se alejan de 1,0 (al multiplicar la probabilidad pre-prueba x 1,0 la probabilidad post-prueba es la misma), y
- cuando los resultados de una prueba se distribuyen en diferentes categorías, los valores extremos de la prueba (los más altos y los más bajos) resultan en RV imprecisas.

Recordemos que:

- Cuando mayor sea la probabilidad pre-prueba, tanto mayor será la probabilidad post-prueba.
- **Una LR > 10 significa que la prueba es adecuada para confirmar un diagnóstico, y < 0,1, que es adecuada para descartarlo.**
- Cuando la probabilidad pre-prueba es muy alta o muy baja, es improbable que las RV cambien la probabilidad post-prueba. Las pruebas son más útiles cuando esta probabilidad pre-prueba se sitúa en torno al 50%.
- Cuando se utilicen pruebas diagnósticas en secuencia, la probabilidad post-prueba de la primera es la probabilidad pre-prueba de la segunda.

4. Utilidad clínica

La utilidad clínica expresa dos conceptos. Primero, el grado en que la validez técnica y analítica que se ha demostrado que tiene una prueba diagnóstica redundan de algún modo y en qué medida o no en los resultados en salud. Por ejemplo, si se acorta el período entre el diagnóstico y el inicio del tratamiento (y sus consecuencias para el paciente); si cambia la decisión terapéutica que se hubiera tomado con pruebas alternativas disponibles; si modifica mejora el pronóstico; si, al tener menos resultados positivos y negativos falsos, se reducen pruebas innecesarias, algunas de las cuales pueden dañar, así como la incertidumbre y la angustia de pacientes y familiares; si se realiza en menor tiempo y es menos invasiva que las pruebas diagnósticas empleadas hasta la fecha y con ello se reduce el riesgo de causar daño asociado con las pruebas adicionales, etc. Segundo, la medida en que los beneficios de la nueva prueba son superiores a los riesgos que acarrea (su relación riesgo/beneficio).

En el siguiente cuadro se ofrecen más ejemplos de validez clínica y utilidad clínica para pruebas con diferente finalidad (diagnóstica, cribado, evaluación del riesgo, pronóstica, de respuesta y toxicidad a medicamentos).

Table 1 Categories of genetic test applications and some characteristics of how clinical validity and utility are assessed

Application of test	Clinical validity	Clinical utility
Diagnosis (symptomatic patient)	Association of marker with disorder	Improved clinical outcomes ^a —health outcomes based on diagnosis and subsequent intervention or treatment Availability of information useful for personal or clinical decision-making End of diagnostic odyssey
Disease screening (asymptomatic patient)	Association of marker with disorder	Improved health outcome based on early intervention for screen positive individuals to identify a disorder for which there is intervention or treatment, or provision of information useful for personal or clinical decision making
Risk assessment/susceptibility	Association of marker with future disorder (consider possible effect of penetrance)	Improved health outcomes based on prevention or early detection strategies
Prognosis of diagnosed disease	Association of marker with natural history benchmarks of the disorder	Improved health outcomes, or outcomes of value to patients, based on changes in patient management
Predicting treatment response or adverse events (pharmacogenomics)	Association of marker with a phenotype/metabolic state that relates to drug efficacy or adverse drug reactions	Improved health outcomes or adherence based on drug selection or dosage

^aClinical outcomes are the net health benefit (benefits and harms) for the patients and/or population in which the test is used.

<https://www.cdc.gov/egapreviews/methods.html>

5. Algunas cautelas

En la evaluación de pruebas diagnósticas, y sobre todo con nuevos biomarcadores y pruebas de diagnóstico genético, cabe resaltar varios hechos, nada infrecuentes:

- **La valoración completa de cualquier prueba incluye la validez técnica y clínica y la utilidad clínica.** Puede ocurrir, y ocurre, que lo que detecta una nueva prueba con alta validez técnica apenas tenga utilidad clínica (por ej., mutaciones sin significado clínico concreto aún conocido). Incluso es posible que pruebas con alta validez técnica y clínica carezcan de utilidad clínica o esta no sea mayor que la de otras que ya se utilizan.
- Es frecuente que pruebas con validez técnica tengan baja validez clínica o que la evaluación de ésta última sea a todas luces insuficiente y no sean autorizadas (por ej., una fracción importante de biomarcadores presentados a las agencias reguladoras). A la postre, **la validez clínica es un parámetro intermedio de lo que realmente interesa, la utilidad clínica**, y es erróneo considerarlo un buen sustituto (*surrogate*) de la utilidad clínica si no se ha demostrado que las variables sustitutivas *predicen o discriminan* bien las de resultado final (para entender el significado de esta última frase, véase la Parte II de este capítulo).
- Por último, la evaluación de una prueba diagnóstica, para ser completa, debe incorporar la valoración de su **valor diagnóstico añadido** (también llamado incremental o relativo, es decir su valor diagnóstico comparado con el de las mejores pruebas diagnósticas utilizadas hasta la fecha para el mismo diagnóstico). Dicho

valor viene dado por las mejoras adicionales que ofrece respecto a esas otras pruebas disponibles en relación con (comparada con):

- primero, sus mayores validez técnica y clínica y su utilidad clínica, y
- segundo, su eficiencia (su relación de coste-efectividad y coste-utilidad incrementales). Por el lado de los costes, por ejemplo, por necesitar aparatos de menor precio, menos personal en laboratorios para realizarla o mucho menos personal aún, por ejemplo, con aquellas tecnologías *point of care* que realmente cumplan los criterios anteriores, o incluso ningún personal salvo el mismo paciente (por ej., glucómetros portátiles, autotoma de tensión), menos mantenimiento, etc. Por el lado de la efectividad, porque mejora su validez clínica o su utilidad clínica. Puede ocurrir que el coste de una nueva prueba sea mayor que el de su alternativa utilizada (coste incremental), con el mismo o mayor valor diagnóstico. Siempre hay que preguntarse si la magnitud del incremento del coste se corresponde con (o justifica) la magnitud del incremento de su efectividad (“si vale lo que cuesta”) o no.

Para acabar esta Capítulo, se ha de estudiar la segunda parte.

6. Material didáctico complementario

Grimes DA, Shulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet*. 2005;365:1500-5.

7. Bibliografía

General

Nivel introductorio

- Gordis L. *Epidemiology*, Second Edition. Philadelphia, PA: WB Saunders Company; 2000.
- Mausner JS, Kramer S. *Epidemiology: An introductory text*. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. *Foundations of epidemiology*. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Epiemiología clínica. Ciencia básica para la medicina*. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. Epidemiology, Beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc.; 2000.

Específica de este Capítulo

- En las referencias del nivel introductorio y del intermedio hay capítulos que tratan estos temas en detalle y con ejemplos.
- Campillo-Artero C. Regaining Health Technology Assessment from Oblivion: Improving and Integrating Regulation of Drugs, Medical Devices, Diagnostic Tests and Surgical Innovations. In: del Llano-Señarís J, Campillo-Artero C, eds. Health technology assessment and health policy today: a multifaceted view of their unstable crossroads. Madrid: Springer Healthcare; 2014:51-73.
- Campillo-Artero C. The regulation of biomarkers. In: Rey-Lopez JC, Lumbreras B, Ponce-Lorenzo JJ, Campillo-Artero C, Maria Pastor-Valero. Precision medicine: Will technology be leveraged to improve population health? Amsterdam: Springer; 2019.

Capítulo 6. Pruebas diagnósticas. Parte II

Índice

1. Introducción	98
2. Capacidad discriminante y construcción de curvas ROC	98
3. Comparación de biomarcadores: capacidad discriminante incremental	102
4. Medidas de asociación, capacidad discriminante y factores de riesgo.....	105
5. Evaluando la calibración y la capacidad discriminante de los modelos multivariantes	105
6. Un último ejempl	112
7. Recapitulación	114
8. Material didáctico complementario.....	115
9. Bibliografía.....	116

Objetivos específicos de este Capítulo

1. Entender el significado de la capacidad discriminante de las pruebas diagnósticas, biomarcadores, factores de riesgo y modelos multivariantes y aprender a estimarla.
2. Conocer su utilidad en la incorporación y en el uso de nuevas pruebas diagnósticas y biomarcadores en la práctica clínica y aprender a evaluar la de los modelos multivariantes.

Nota importante

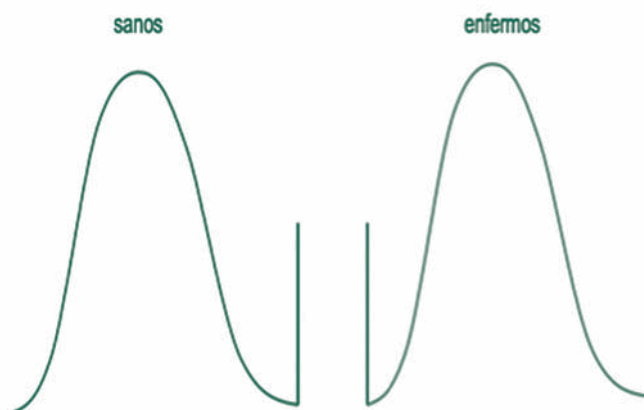
En este Cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I, 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como **Material didáctico complementario**.

1. Introducción

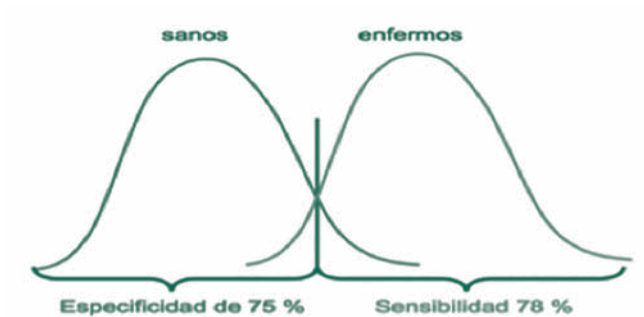
No es propósito de estos dos capítulos profundizar excesivamente en las pruebas diagnósticas. Pero ello no es óbice para hacer una incursión a los objetivos, características y utilidad del estudio de la **capacidad discriminante** (*discriminant accuracy*) de variables, supuestos factores de riesgo, biomarcadores, pruebas diagnósticas y modelos multivariantes mediante las curvas ROC, sobre todo por su importancia en clínica y en investigación. Además, porque con el progreso de la medicina estratificada y personalizada necesitamos entender la validez y utilidad de los nuevos biomarcadores. Y, por último, porque todo ello nos ayudará a acabar de entender y a complementar varios conceptos importantes relacionados con la asociación, el riesgo, la causalidad, así como a conocer los errores que con frecuencia se cometen al interpretarlos.

2. Capacidad discriminante y construcción curvas ROC

Muchas pruebas diagnósticas solo tienen dos resultados posibles: positivo (tiene la enfermedad o el factor de riesgo o cierta característica) o negativo (no la tiene), que se definen, la mayoría de las veces, por consenso fijando un punto de corte (*cut-off point*). Pero es muy infrecuente que un punto de corte discrimine sin error (sensibilidad y especificidad = 100%) a los enfermos y a los sanos, a los que desarrollan o no cierta enfermedad, responden o no a un medicamento, tienen cierto atributo o no, si un medicamento será o no tóxico. La siguiente figura ilustra esta inusual situación.

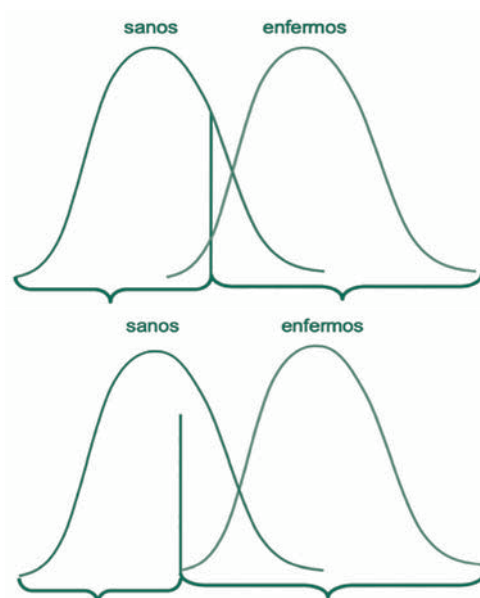


Lo más habitual es encontrarnos en la situación en que para un punto de corte seleccionado, con mayor o menor acierto, algunos sanos se diagnosticarán como enfermos (falsos positivos, menor especificidad) y algunos enfermos se considerarán sanos (falsos negativos, baja sensibilidad). Este caso, mucho más frecuente, se aprecia en la siguiente figura.



¿Qué diferencia real existe entre una persona con un valor de la prueba una unidad por encima del punto de corte y otra con un valor una unidad por debajo?

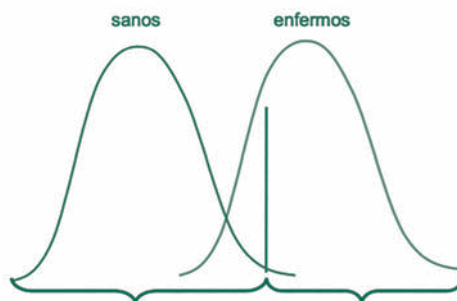
Es importante averiguar cómo varían los parámetros de validez clínica de una prueba diagnóstica con diferentes puntos de corte (aunque parezca algo engorroso lo que sigue, es importante y merece la pena leerlo pausadamente):



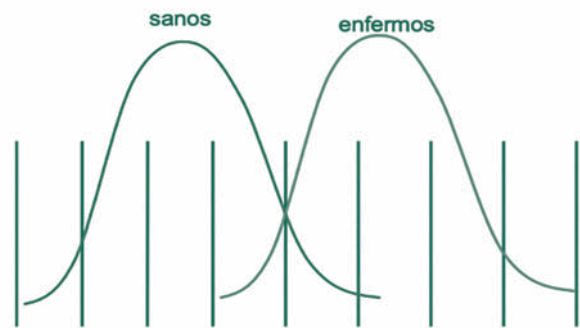
Si desplazamos el punto de corte hacia la izquierda, como en esta figura, más sanos se diagnostican como enfermos (más falsos positivos, menos especificidad) y menos enfermos como sanos (menos falsos negativos, más sensibilidad).

Incluso si fijamos un punto de corte con ningún falso negativo (máxima sensibilidad), lo hacemos a expensas de perder especificidad (aumentan los falsos positivos, más sanos se consideran enfermos).

Por el contrario, si desplazamos el punto de corte hacia la derecha, se produce el efecto contrario: se reducen los falsos positivos (aumenta la especificidad) y aumentan los falsos negativos (disminuye la sensibilidad, más enfermos se consideran sanos).

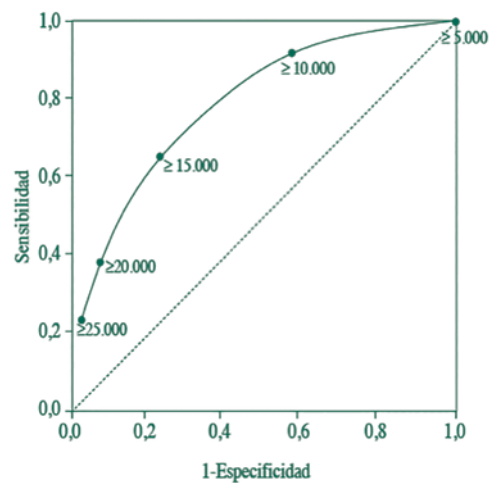


Y esto puede analizarse con cuantos puntos de corte creamos necesarios. Es lo que han hecho los investigadores de un estudio sobre la validez clínica del recuento de leucocitos en la sangre para diagnosticar bacteriemia en niños seleccionando los cinco puntos de corte (recuento de leucocitos/mm³) que aparecen en las siguientes figura y tabla (fijémonos cómo varían los parámetros de validez clínica con cada uno de los cinco puntos de corte):



Recuento de leucocitos / mm ³	VP	FP	VN	FN	Se	Es	1-Es
≥ 5.000	26	859	0	0	1,00	0,00	1,00
≥ 10.000	24	493	366	2	0,92	0,43	0,57
≥ 15.000	17	200	659	9	0,65	0,77	0,23
≥ 20.000	10	70	789	16	0,38	0,92	0,08
≥ 25.000	6	26	833	20	0,23	0,97	0,03

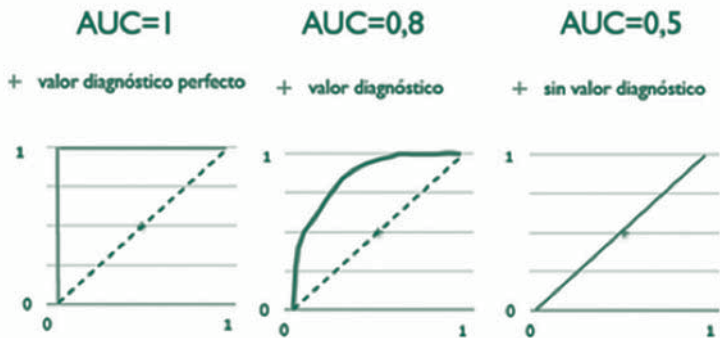
Cuando los datos de la tabla anterior se representan gráficamente en una curva con las cifras de la sensibilidad (% de verdaderos positivos) en el eje de las Y y 1 – especificidad (% de positivos falsos) en el eje de las X, se obtiene una curva, llamada ROC, como la siguiente:



<https://www.aeped.es/sites/default/files/anales/50-3-19.pdf>

Esta curva ROC representa la relación entre los verdaderos positivos (eje Y, sensibilidad) y los falsos positivos (eje X, 1 - especificidad) e informa sobre la capacidad de la prueba, la variable, el supuesto factor de riesgo, el biomarcador o de un modelo predictivo de discriminar (discernir, separar, diferenciar, predecir) los que tienen y no tienen cierta característica (por ej., enfermedad). Veamos cómo se interpreta.

Con programas estadísticos se calcula el área bajo la curva (ABC, *area under the curve*, AUC) de cada prueba. Cuanto más próxima a 1,0 es, mayor ABC y mayor capacidad discriminante tiene, porque con sensibilidad 100%, el % de falsos negativos = 0 y con especificidad = 100%, (1-100 en el eje X del gráfico) el de falsos positivos = 0. Un valor 0,5 indica (diagonal) que no tiene capacidad discriminante, porque en este punto, con sensibilidad = 50% (50% de falsos negativos) y especificidad de 50% (50% de falsos positivos) la prueba no discrimina mejor que el azar, que tirar una moneda al aire. El ABC también se conoce como estadístico C. (El valor de este estadístico, junto con su IC95%, puede pedirse a los programas con que se construyen modelos multivariantes cuando queremos conocer su capacidad discriminante. Se comentará un poco más adelante.)

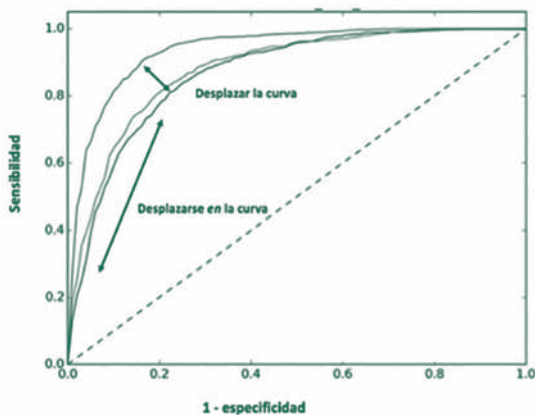


Cuando para un mismo diagnóstico disponemos de varios puntos de corte, cada uno de ellos con su sensibilidad y especificidad correspondientes, escogeremos aquel que maximiza los valores de estos dos parámetros simultáneamente, 0,78 en el caso del estudio cuyos resultados sobre el recuento de leucocitos aparecen a continuación. Con este punto de corte la prueba alcanza su máximo rendimiento (su sensibilidad y especificidad son las más altas conjuntamente.

Resultado del test mg/ml	Sensibilidad	Especificidad	
• 12 =	100 %	20 %	
• 23 =	98 %	34 %	
• 35 =	90 %	50 %	
• 45 =	84 %	62 %	
• 78 =	78 %	75 %	mejor rendimiento
• 99 =	47 %	89 %	
• 120 =	10 %	100 %	

Reparemos en un detalle: al hacer un ejercicio como el anterior de la leucocitosis y comprobar que, modificando el punto de corte, cuando aumenta la sensibilidad, descende la especificidad, y viceversa, lo que estamos haciendo es *desplazarnos* en la curva ROC (*moving along the curve*), como hacemos cuando analizamos en una curva de demanda o de oferta y vemos cómo varía la demanda o la oferta cuando se modifica el precio.

Sin embargo, en la siguiente figura se observa que las ABC de unos biomarcadores son mayores que las de otros (los biomarcadores con mayor área tienen más capacidad discriminante). Lo que realmente innova e interesa es aumentar el ABC de un nuevo biomarcador respecto a los existentes, es decir, alejar la curva ROC de la diagonal en sentido sudoeste y noroeste. **Con la innovación real la curva ROC se desplaza (*shifting the curve*), no nos desplazamos en ella, lo que significa aumentar la capacidad discriminante de los biomarcadores o de los factores de riesgo supuestos que estemos estudiando.** Es lo mismo que *desplazar* la curva de la oferta o la de la demanda al investigar cómo cambia la curva por efecto de variables distintas del precio.



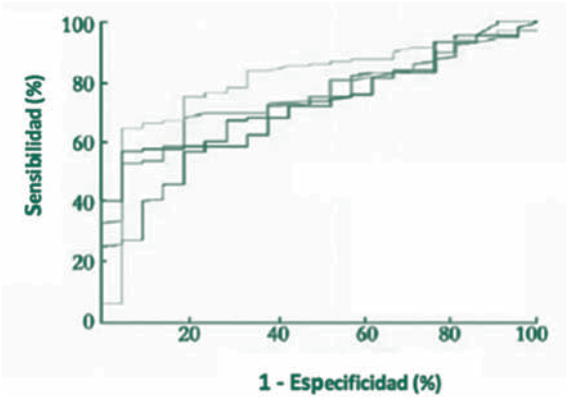
3. Comparación de biomarcadores: capacidad discriminante incremental

Si tenemos más de una prueba, hemos de escoger aquella con mayor capacidad discriminante, es decir la que para los puntos de corte elegidos tenga mayor % de verdaderos positivos (sensibilidad) y menor % de falsos positivos (especificidad) o, en otras palabras, mayor ABC. El siguiente ejemplo se encuentra en este artículo: [Fayyad-Kazan J, et al. Circulating miR-150 and miR-342 in plasma are novel potential biomarkers for acute myeloid leukemia. J Translational Med. 2013;11:31.](#)

En él aparecen las curvas ROC estimadas de dos biomarcadores potenciales y de su uso combinado en la leucemia mieloide aguda. Las estimaciones puntuales de las tres ABC difieren apenas unas centésimas entre sí y los valores de sus IC95% se solapan en gran medida. Por consiguiente, sus capacidades discriminantes usados individualmente y en combinación son casi iguales o iguales. En el estudio no se notifican los IC95% de la sensibilidad y la especificidad, aunque cabe esperar que buena parte de sus valores también

se solapan. No obstante, debe tenerse en cuenta el carácter preliminar de los resultados de este estudio, probablemente, el primero publicado sobre estos biomarcadores, que deberán confirmarse en estudios posteriores.

Este otro ejemplo muestra las ABC de cuatro marcadores de presencia de un tipo de cáncer.



No cabe duda de que deberíamos escoger el de la amarilla, porque su ABC es la mayor y, por ello, su capacidad discriminante (es incremental) es la mayor de los cuatro marcadores, aunque como en el caso anterior, las ABC no se diferencian en gran medida y, aunque los autores nos los notifiquen, los IC95% también han de compartir muchos valores. (Más adelante se hacen recomendaciones sobre la selección de marcadores a partir de su capacidad discriminante incremental.)

Si tenemos más de una prueba, hemos de escoger aquella con mayor capacidad discriminante, es decir la que para los puntos de corte elegidos tenga mayor % de verdaderos positivos (sensibilidad) y menor % de falsos positivos (especificidad) o, en otras palabras, mayor ABC.

En las siguientes tablas se presentan los resultados de un estudio para conocer la capacidad discriminante de los supuestos factores de riesgo tradicionales de enfermedad coronaria. Apréciase que las estimaciones puntuales de las ABC se acompañan de sus IC95%, como debería ser en todos los artículos publicados.

	AU-ROC (95% CI)	
		Difference
Age	0.63 (0.60 – 0.65)	-0.05
Sex*	0.61 (0.59 – 0.64)	-0.07
Age and sex	0.68 (0.66 – 0.70)	Reference
Traditional risk factors		
- Systolic blood pressure*	0.65 (0.62 – 0.67)	-0.03
- Diastolic blood pressure	0.60 (0.58 – 0.62)	-0.08
- Hypertension arterial	0.58 (0.56 – 0.61)	-0.10
- Glucose*	0.61 (0.59 – 0.64)	-0.07
- Diabetes	0.57 (0.54 – 0.59)	-0.11
- Total cholesterol	0.57 (0.54 – 0.59)	-0.11
- HDL cholesterol	0.64 (0.61 – 0.66)	-0.04
- LDL cholesterol	0.58 (0.55 – 0.61)	-0.10
- LDL/HDL ratio*	0.65 (0.63 – 0.68)	-0.03
- Triglycerides	0.61 (0.58 – 0.63)	-0.07
- Body Mass Index*	0.59 (0.56 – 0.61)	-0.09
- Cigarette smoking*	0.58 (0.55 – 0.60)	-0.14

Merlo J, Wagner P, Juarez S, Mulinari S, Hedblad B. The tyranny of the averages and the indiscriminate use of risk factors and population attributable fractions in Public Health: the case of coronary heart disease. Sweden Malmö, Sweden: Unit for Social Epidemiology, Department of Clinical Sciences, Faculty of Medicine, Lund University; 2013. (Working paper version 2013.09.26.) Acceso libre.

La columna de la derecha indica la diferencia de la capacidad discriminante de cada supuesto factor de riesgo (FR) respecto a la edad y el sexo juntos, que, como se observa, es menor para todos los FR tradicionales. En otras palabras, todos los FR individuales tienen menor capacidad discriminante que la edad y el sexo juntos.

Y lo mismo ocurre con los biomarcadores de enfermedad coronaria, también en relación con el sexo y la edad juntos:

Biomarkers		
– CRP	0.61 (0.58 – 0.63)	-0.07
– Cystatin C	0.62 (0.59 – 0.65)	-0.06
– Lp-PLA2 activity	0.61 (0.58 – 0.64)	-0.07
– N-BNP	0.54 (0.51 – 0.57)	-0.14
– N-BNP (309 ng/L)	0.52 (0.49 – 0.55)	-0.16

Por último, en el siguiente cuadro se presentan las ABC y los cambios en la capacidad discriminante incremental (respecto a la de la edad y el sexo juntos) de combinaciones de factores de riesgo, biomarcadores y la combinación de edad + sexo.

Combinations		
– Traditional risk factors	0.71 (0.69 – 0.74)	0.03
– Biomarkers	0.67 (0.64 – 0.69)	-0.01
– Biomarkers and traditional	0.74 (0.71 – 0.76)	0.06
– Age, sex and traditional risk factors	0.75 (0.73 – 0.77)	0.07
– Age, sex and biomarkers	0.72 (0.69 – 0.74)	0.04
– Age, sex, traditional risk factors and biomarkers	0.76 (0.74 – 0.79)	0.08

La capacidad discriminante del efecto combinado de todos los supuestos FR sólo es 0,03 veces más alta que la de la edad + sexo. La de todos los biomarcadores juntos, 0,01 menor, y la de los FR y biomarcadores juntos, 0,06 mayor, siempre en relación con lo que aumentan (por añadidura) o no la capacidad discriminante de la edad y el sexo juntos. Se comprueba finalmente que el máximo incremento de la capacidad discriminante se obtiene con la utilización conjunta de (edad + sexo) + FR + biomarcadores, aunque sólo es 0,08 unidades (casi un 12%) más elevada que la de la edad y el sexo juntos.

Comparando el precio de un nuevo biomarcador con el de los utilizados podemos saber si el nuevo “vale lo que cuesta”, si su capacidad discriminante incremental justifica su aumento precio, como hacemos con los nuevos medicamentos o aparatos médicos. Conociendo el precio y el número de personas potenciales a las cuales se pedirían todas estas pruebas, por ejemplo, en un año, podemos decidir si usar simultáneamente todos los marcadores vale (valor diagnóstico incremental) lo que cuesta y aproximarnos a su impacto presupuestario (el gasto de todas las pruebas = precio x cantidad).

4. Medidas de asociación, capacidad discriminante y factores de riesgo

Es muy importante dejar claro que en muchísimos estudios se olvida que **las medidas de asociación medias (OR, RR, HR) entre supuestos factores de riesgo o biomarcadores (diagnósticos, pronósticos, de respuesta, toxicidad o predictivos) y variables de resultado o de respuesta son inadecuadas para evaluar su rendimiento.** Las fuerzas de asociación fuertes (OR, RR o HR > 10) se relacionan con una baja capacidad de los supuestos factores de riesgo y de los biomarcadores para discriminar entre casos y no casos en la población general. Incluso las exposiciones (por ej., factores de riesgo utilizados como pruebas de cribado, mutaciones genéticas, signos y síntomas, etc.) de muy alto riesgo de enfermedad o de resultado tienen una baja prevalencia en la población general y explican una fracción pequeña de casos. Esta es la razón por la cual las medidas de asociación media tienen capacidad discriminante tan baja y han de interpretarse, por consiguiente, junto con medidas de su capacidad discriminante, lo que, como se ha indicado, se olvida con mucha frecuencia.

Revisemos este artículo: [Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004; 159:882–90.](#) Como indican las curvas ROC estimadas, para alcanzar una sensibilidad del 85% y una especificidad del 85% $(1 - \text{especificidad}) = (1 - 85) = 0,15$ en el eje de las x, necesitaríamos un OR igual a, aproximadamente, 75 (con alta capacidad discriminante, pero casi nunca visto). Dicho de otro modo, con un OR = 3 tendríamos una sensibilidad aproximadamente del 20% (muy baja, muchos falsos negativos) y una especificidad del 90% (alta, con un 10% de falsos positivos), es decir una capacidad discriminante, estimada con el ABC cercana al 60%, un poco más alta que la del azar o que echar una moneda al aire $\approx 50\%$, la diagonal. Este es un artículo sumamente importante, más aún para entender la era de la medicina de precisión.

Deben subrayarse dos hechos muy importantes y sus peligros: 1) el abuso de atribuciones de naturaleza causal entre exposiciones y sus efectos solamente sobre la base de medidas de la fuerza de asociación, y 2) considerar como factor de riesgo a una variable solamente porque, a juzgar por el OR o el RR, está asociada con un efecto sin haber comprobado su capacidad discriminante.

5. Evaluando la calibración y la capacidad discriminante de modelos multivariantes

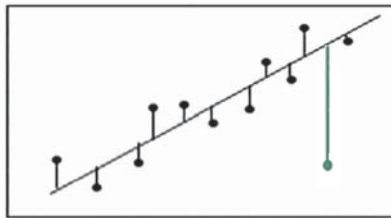
Recordemos dos conceptos del análisis estadístico, que, no por omitidos con harta frecuencia en los estudios publicados, dejan de ser esenciales. Cuando se evalúan modelos estadísticos multivariantes publicados en revistas (o aquellos con que estemos trabajando), antes de creernos las interpretaciones que hacen los autores de los resultados o de

estar seguros de las nuestras, es extremadamente importante saber, dicho mal y pronto, *si el modelo funciona*. Esto se circunscribe a conocer dos cosas, la calibración o bondad de ajuste del modelo y su capacidad discriminante:

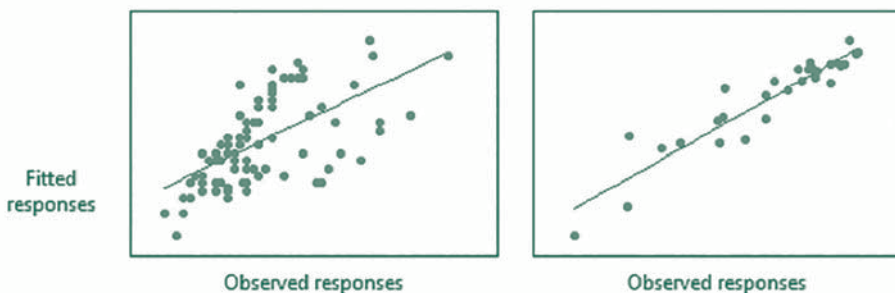
- 1) la **calibración o bondad de ajuste del modelo** (*calibration, goodness-of-fit*), esto es, la medida en que el modelo se ajusta bien a los datos observados, es decir, el grado en que los datos *esperados* calculados por el modelo coinciden con los *observados*. Este concepto se ha descrito en detalle en el cuaderno de estadística y
- 2) la capacidad discriminante (*discriminant accuracy*) del modelo, que viene expresada por el ABC o estadístico C.

Ilustremos con unos ejemplos gráficos el significado (intuitivo) de la calibración o bondad de ajuste de un modelo.

En esta primera figura se muestra la recta de regresión que ha creado el modelo, la recta que minimiza las diferencias entre los valores observados por los investigadores (los medidos, los puntos encima y debajo de la recta) y los valores *estimados* o *ajustados* por el modelo (también conocidos como esperados, los que están en la recta que construye el modelo), junto con los denominados **residuales**, que son igual a la diferencia entre los valores observados (puntos negros) y los esperados (que se encuentran exactamente en la línea recta). Se aprecia que unos residuales son mayores que otros (la distancia a la recta de cada uno de ellos varía, y es máxima, el residual de mayor valor, en el valor observado verde):

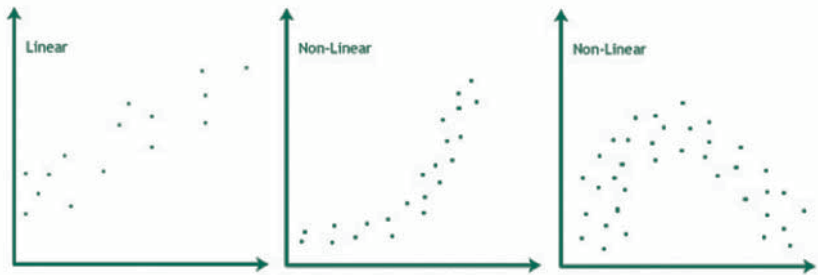


Las dos figuras de abajo muestran claramente que la bondad de ajuste del modelo de la izquierda es mucho peor que la del modelo de la derecha, porque, sencillamente, los residuales de la izquierda son mayores que los de la derecha (su distancia a la línea recta es mayor, la diferencia entre los valores observados y los estimados o ajustados por el modelo lo es). En el de la derecha bastantes puntos se acercan mucho más a la recta o se encuentran en ella. Si el ajuste de un modelo fuese perfecto, todos los puntos se encontrarían en la línea recta.



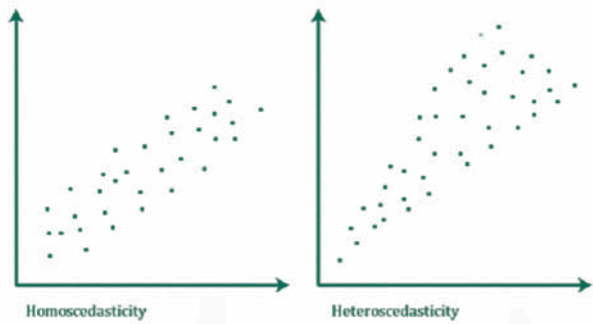
El conjunto de los residuales (lo que no ajusta el modelo, que se denomina error, y es el último término (e) que aparece en las fórmulas de los modelos) es la parte de la variable dependiente que el modelo, con las variables independientes que hemos decidido o podido incluir, no es capaz de explicar o predecir, según sea el objetivo con que se ha construido el modelo.

Siempre hay que ver estas gráficas de los residuales, porque son muy informativas, por añadidura a los resultados estadísticos de los modelos. Sin entrar en detalles ni complejidades, fijémonos en las siguientes dos series de representación gráfica de los residuales.



Vemos claramente en esta primera serie que los valores observados (los reales que se han medido) tienen distribuciones diferentes (una lineal y la las otras dos no). Si el modelo trata la del centro y la de la derecha como lineales, por mucha parafernalia estadística que veamos, no puede ajustarse bien a esas distribuciones con una línea recta. Los resultados que dé (los coeficientes de regresión, y las medidas de asociación que se calculan a partir de ellos, en las dos estarán sesgados. Solo podría hacerlo bien con la de la izquierda si contiene variables predictivas o explicativas suficientes, lo que depende del conocimiento que tengamos y de la disponibilidad de datos.

En la siguiente serie vemos que la distribución de los residuales respecto a la línea de regresión dada por el modelo es distinta. En la de la izquierda, la distribución es homogénea (los residuales se distribuyen encima y debajo de la recta de forma similar y homogénea, las distancias de los valores observados a la recta lo son, es decir, la variación o varianza de los residuales es más o menos la misma alrededor de toda la recta). Por el contrario, en la de la derecha, la distancia aumenta hacia la derecha de la recta; la varianza de los residuales aumenta, no es homogénea ni constante en todo el recorrido de la recta de regresión construida por el modelo con los valores esperados.



¿Y qué importa todo esto? Importa mucho. En una fracción sustancial de modelos publicados se supone que los errores, los residuales, se distribuyen de forma aleatoria y homogénea como los de la figura de la izquierda (hay, como se dice en este caso, homocedasticidad) y su comprobación se olvida o se pasa por alto. Si se supone esto y los residuales son los de la figura de la derecha (incumplen el principio de distribuirse aleatoriamente de forma homogénea, hay heterocedasticidad), los resultados del modelo también están sesgados.

Veamos la utilidad de lo que acabamos de explicar con la tabla I, incluida en el artículo (también citado en el Capítulo 6) [Garbe C, et al. Risk Factors for Developing Cutaneous Melanoma and Criteria for Identifying Persons at Risk: Multicenter Case-Control Study of the Central Malignant Melanoma Registry of the German Dermatological Society. Invest Dermatol. 1994;102:695-9.](#)

En esta tabla aparecen los resultados de una regresión logística con los *supuestos* factores de riesgo del melanoma cutáneo (CM). Basándose en ella, los investigadores concluyen que los seis barajados son factores de riesgo independientes del CM. En este artículo no se notifica ni la bondad de ajuste del modelo ni su capacidad para discriminar qué personas con dichos supuestos factores de riesgo desarrollan o no melanoma cutáneo. Sin esta información, ni podemos conocer la capacidad discriminante de cada uno de ellos individualmente ni la de sus diferentes combinaciones hasta la de los seis en conjunto, que sería el ABC del modelo logístico presentado aquí. Tampoco sabemos si el modelo tiene una bondad de ajuste aceptable. Por consiguiente, careciendo de esta información, en rigor, no se puede afirmar que son factores de riesgo independientes de melanoma, como se hace tan a menudo en tantos estudios. (Se trata de otro incumplimiento habitual de muchas revistas médicas de las normas de publicación consensuadas a escala internacional.)

La tabla que aparece a continuación corresponde a un estudio sobre *fallo en el rescate* (*failure to rescue*) de pacientes ingresados en una UCI tras ser sometidos a cirugía cardíaca ([Riera M, Amézaga R, Molina M, Campillo-Artero C, Sáez de Ibarra JI, Bonnín O, Ibáñez J. Mortalidad de las complicaciones postoperatorias \(failure to rescue\) tras cirugía cardíaca en un hospital universitario. Rev Cal Asist.2016;31:126-33\).](#)

Las variables de este modelo final de regresión logística son las asociadas con una o más de nueve complicaciones seleccionadas a priori que aparecen durante y hasta 30 días después de esta cirugía en una cohorte de 2750 pacientes. El objetivo del estudio no es explicativo, sino predictivo. Todas ellas están asociadas con el efecto (aparición de una o más complicación), aunque la fuerza de las asociaciones de cada variable supuestamente predictiva de una o más complicaciones post-cirugía es baja a juzgar por la magnitud de los OR y sus IC95%. Hay asociación, sin duda, el tamaño de la muestra es notable, pero sin más información podríamos extraer conclusiones erradas como esta de nuevo: las seis variables independientes del modelo son factores de riesgo o predictivos de las complicaciones de la cirugía cardíaca escogidas.

Esta conclusión, no es veraz, como tampoco lo es la del caso anterior, y cambia al conocer la capacidad discriminante de este modelo (de las variables incluidas en el modelo tomadas en conjunto) para identificar qué pacientes hacen complicaciones y qué pacientes no las hacen. Según la prueba de bondad de ajuste (de Hosmer-Lomeshow), el modelo tiene buena calibración, ajusta razonablemente bien, pero como indica el ABC o estadístico C del modelo (0,66; IC95%: 0,64-0,67), su capacidad discriminante con estas seis variables es baja para discriminar o predecir qué pacientes intervenidos tendrán una o más de dichas complicaciones.

Además, como traduce este ejemplo, es muy raro que una variable (supuesto factor de riesgo o un signo o síntoma) por sí sola permita discriminar quién desarrollará una complicación, quién tiene una enfermedad determinada porque todos sus RV+ son menores de 4,0, como se aprecia con frecuencia. Veamos, para insistir y consolidar estos conceptos esenciales, dos ejemplos más.

Durante décadas, los signos de Kernig y Brudzinski se han interpretado como pruebas de inflamación meníngea bacteriana o tuberculosa sin que se haya estimado su validez clínica completa —incluida su capacidad discriminante— (los tienen una sensibilidad de 57 y 96%, respectivamente). Evalúese su capacidad para confirmar la inflamación meníngea a nivel individual a partir de las RV+ y las probabilidades post-prueba que aparecen en la tabla 2 de este artículo: [Thomas KE, et al. The Diagnostic Accuracy of Kernig's Sign, Brudzinski's Sign, and Nuchal Rigidity in Adults with Suspected Meningitis. Clin Inf Dis. 2002;35:46-52.](#)

Y todos estos resultados son lógicos si reparamos en que **la mayoría de los diagnósticos los establecemos combinando varios signos (incluidos los resultados de diversas pruebas diagnósticas) y síntomas y otras pruebas complementarias de laboratorio, no con uno solo.**

El siguiente ejemplo sobre factores de riesgo de preeclampsia es muy ilustrativo, porque el estudio incluye toda la información necesaria para conocer la capacidad discriminante de los factores considerados tradicionalmente de riesgo de la preeclampsia. Incluye, además, las razones de verosimilitud. Ninguna RV+ es $> 5,0$ y ninguna RV – es $< 0,8$, valores muy lejanos a los puntos de corte mencionados en la parte I del Capítulo 6 (> 10 y $< 0,1$), respectivamente, para poder confirmar o descartar razonablemente, en este caso, preeclampsia.

Table 3. Relation between sociodemographic, pre-pregnancy and pregnancy characteristics and risk of preeclampsia (PE) in the 282599 Swedish deliveries from nulliparous woman recorded between 2002 and 2010.

Risk Factor	AR	AF	TPF (95% CI)	FPF (95% CI)	LR+ (95% IC)	LR- (95% CI)	Orc (95% CI)	ORa (95% CI)
Maternal age, y								
<40	5.51						Ref	Ref
≥ 40	9.14	3.63	2.7 (2.4–2.9)	0.6 (0.5–0.6)	1.71 (1.55–1.88)	0.99 (0.99–0.99)	1.73 (1.56–1.91)	1.49 (1.31–1.62)
Educational status								
Medium-High	5.06						Ref	Ref
Low	5.97	0.91	59.6 (58.8–60.3)	55.3 (55.1–55.5)	1.08 (1.06–1.09)	0.90 (0.89–0.92)	1.19 (1.15–1.23)	1.13 (1.09–1.17)
Family situation								
Cohabiting	5.58						Ref	Ref
Non-Cohabiting	5.36	-0.22	5.8 (5.4–6.2)	6 (5.9–6.1)	0.96 (0.90–1.02)	1	0.96 (0.89–1.02)	0.94 (0.88–1.01)
Diabetes								
No	5.46						Ref	Ref
Yes	22.7	17.3	2.5 (2.2–2.7)	0.5 (0.5–0.5)	4.98 (4.45–5.58)	0.98 (0.98–0.98)	5.08 (4.53–5.70)	4.29 (3.77–4.88)
HBP								
No	5.53						Ref	Ref
Yes	24.1	18.5	3.2 (2.9–3.5)	0.6 (0.6–0.6)	5.37 (4.86–5.93)	0.97 (0.97–0.98)	5.51 (4.98–6.10)	4.07 (3.63–4.57)
Autoimmune								
No	5.56						Ref	Ref
Yes	9.27	3.71	0.4 (0.3–0.5)	0.2 (0.2–0.2)	1.73 (1.32–2.27)	1	1.73 (1.32–2.28)	1.12 (0.78–1.60)
CKD								
No	5.55						Ref	Ref
Yes	10.1	4.50	0.9 (0.7–1)	0.4 (0.4–0.5)	1.90 (1.59–2.26)	1 (0.99–1)	1.90 (1.59–2.28)	1.34 (1.09–1.66)
Obesity								
No	4.84						Ref	Ref
Yes	12.1	7.28	21.6 (21–22.3)	9.2 (9.1–9.4)	2.34 (2.27–2.42)	0.86 (0.86–0.87)	2.71 (2.60–2.82)	2.56 (2.45–2.66)
Smoker								
No	5.68						Ref	Ref
Yes	5.13	-0.55	19.3 (18.7–19.9)	21 (20.9–21.2)	0.92 (0.89–0.95)	1.02 (1.01–1.03)	0.89 (0.86–0.93)	0.83 (0.79–0.88)
Multiple pregnancy								
No	5.39						Ref	Ref
Yes	19.7	14.3	4.3 (4–4.7)	1 (1–1.1)	4.17 (3.84–4.52)	0.97 (0.96–0.97)	4.31 (3.96–4.69)	4.27 (3.90–4.67)
Conception by ART								
No	5.48						Ref	Ref
Yes	7.59	2.11	5.4 (5–5.7)	3.8 (3.8–3.9)	1.39 (1.30–1.49)	0.98 (0.98–0.99)	1.41 (1.32–1.52)	1.17 (1.08–1.25)
Gestational diabetes								
No	5.51						Ref	Ref
Yes	13.9	8.40	1.8 (1.6–2)	0.7 (0.6–0.7)	2.74 (2.42–3.10)	0.99 (0.99–0.99)	2.77 (2.44–3.14)	2.04 (1.79–2.33)

AF: Attributable fraction, AR: Attributable risk, TPF: True positive fraction, Orc: crude Odds ratio; ORa: mutually adjusted Odds ratios, LR+: Positive Likelihood ratio, LR-: Negative Likelihood ratio, ART: assisted reproductive technology, CKD: Chronic kidney disease, HBP: Chronic hypertension.

Rodriguez-Lopez M, Wagner P, Perez-Vicente R, Crispi F, Merlo J. Revisiting the discriminatory accuracy of traditional risk factors in preeclampsia screening. PLoS ONE. <https://doi.org/10.1371/journal.pone.0178528>. Acceso libre.

Sí alcanzan $RV+ > 10,0$ y $LR- < 0,1$ algunas de las combinaciones de estos supuestos factores de riesgo, como muestra el siguiente cuadro extraído del mismo artículo que el anterior:

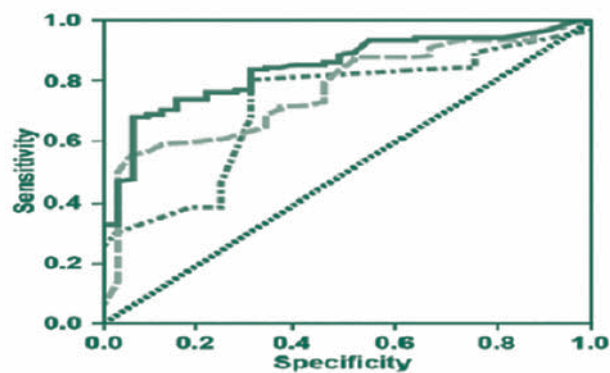
Table 5. Discriminatory accuracy for specific combinations of risk factor for preeclampsia based on random forest variable importance.

Combination				n	AR	AF	TPF (95% CI)	FPF (95% CI)	LR+ (95% CI)	LR- (95% CI)	ORa (95% CI)
Multiparous Incidence PE 2.41%											
PP	HBP	Obese	Multiple	283462	1.23	Ref					
0	0	0	1	4698	8.56	7.33	4.8 (4.4–5)	1.3 (1.2–1.3)	3.79 (3.43–4.19)	0.96 (0.96–0.97)	7.02 (6.28–7.83)
0	0	1	0	39533	3.79	2.56	18.1 (17.3–18.9)	11.3 (11.2–11.4)	1.60 (1.25–1.62)	0.92 (0.91–0.93)	3.05 (2.86–3.25)
0	0	1	1	708	13	11.8	1.1 (0.9–1.4)	0.2 (0.2–0.2)	6.05 (4.86–7.52)	0.99 (0.99–0.99)	11.1 (8.86–13.8)
0	1	0	0	1453	14.7	13.5	2.6 (2.2–2.9)	0.4 (0.3–0.4)	6.96 (6.02–8.03)	0.98 (0.97–0.98)	11.6 (9.94–13.6)
0	1	0	1	24	16.7	15.5	~0	~0	8.10 (2.77–23.7)	~1	14.6 (4.86–43.9)
0	1	1	0	706	19.6	18.4	1.7 (1.4–2)	0.2 (0.1–0.2)	9.84 (8.18–11.8)	0.99 (0.98–0.99)	16.2 (13.2–19.8)
0	1	1	1	10	50	48.8	0.1 (0–0.1)	~0	40.5 (11.7–139)	~1	57.2 (9.99–328)
1	0	0	0	9635	15.1	13.9	17.5 (16.7–18.4)	2.4 (2.4–2.5)	7.20 (6.84–7.58)	0.85 (0.84–0.85)	14 (13.2–15.2)
1	0	0	1	136	34.6	33.4	0.6 (0.4–0.6)	~0	21.4 (15–30.4)	0.99 (0.99–1)	41 (28.7–58.8)
1	0	1	0	2935	24.8	23.6	8.8 (8.2–9.4)	0.3 (0.2–0.3)	13.4 (12.3–14.5)	0.92 (0.91–0.92)	25.8 (23.5–28.3)
1	0	1	1	43	39.5	38.3	0.2 (0.1–0.3)	~0	26.5 (14.4–48.8)	~1	51.9 (28–96.5)
1	1	0	0	410	31.5	30.3	1.6 (1.3–1.8)	0.1 (0.1–0.1)	18.6 (11.6–22.9)	0.99 (0.98–0.99)	32 (26.2–40.6)
1	1	0	1	6	16.7	15.5	~0	~0	8.10 (0.95–69.3)	~1	13.9 (1.8–105)
1	1	1	0	242	33.8	32.6	1 (0.8–1.2)	~0	20.8 (15.9–27.1)	0.99 (0.99–0.99)	36.1 (27.6–40.6)
Multiple model including single variables							AUC:75.35(74.74–75.97) Sensitivity (FP10%):45.01(43.87–46.15) AIC:67934				
Multiple model including combination of variables							AUC:75.23(74.61–75.85) Sensitivity (FP10%):43.34(42.23–44.44) AIC:67632				
Primiparous Incidence PE = 5.57%											
DM	HBP	Obese	Multiple	248600	4.47	Ref					
0	0	0	1	3056	19.1	14.6	3.7 (3.4–4)	0.9 (0.9–1)	4 (3.66–4.37)	0.97 (0.97–0.97)	4.84 (4.14–5.31)
0	0	1	0	26939	11.5	7.03	19.6 (19–20.2)	8.9 (8.8–9)	2.19 (2.12–2.27)	0.88 (0.88–0.89)	2.69 (2.59–2.81)
0	0	1	1	358	23.5	19	0.5 (0.4–0.5)	0.1 (0.1–0.1)	5.20 (4.07–6.64)	1 (0.99–1)	6.10 (4.76–7.84)
0	1	0	0	1418	22.5	18	2 (1.8–2.2)	0.4 (0.4–0.4)	4.88 (4.31–5.53)	0.98 (0.98–0.99)	5.87 (5.16–6.67)
0	1	0	1	17	47.1	42.6	0.1 (0–0.1)	~0	15.1 (5.82–39)	~1	17.4 (6.63–45.5)
0	1	1	0	505	29.7	25.2	1 (0.8–1.1)	0.1 (0.1–0.11)	7.17 (5.93–8.67)	0.99 (0.99–0.99)	8.15 (6.70–9.91)
0	1	1	1	6	16.7	12.2	~0	~0	3.39 (0.40–29)	~1	3.74 (0.38–36.4)
1	0	0	0	1316	21.7	17.2	1.8 (1.6–2)	0.4 (0.4–0.4)	4.71 (4.14–5.36)	0.99 (0.99–0.99)	5.91 (5.18–6.76)
1	0	0	1	13	23.1	18.6	~0	~0	5.09 (1.40–18.5)	~1	6.03 (1.60–22.6)
1	0	1	0	237	30.4	25.9	0.5 (0.4–0.6)	0.1 (0.1–0.1)	7.40 (5.61–9.76)	~1	8.99 (6.80–11.9)
1	0	1	1	4	50	45.5	~0	~0	16.96 (2.39–120)	~1	18.9 (2.98–120)
1	1	0	0	93	12.9	8.43	0.1 (0–0.1)	~0	2.51 (1.37–4.61)	~1	2.42 (1.24–4.69)
1	1	1	0	37	29.7	25.2	0.1 (0–0.1)	~0	7.18 (3.55–14.5)	~1	7.84 (3.84–16)
Multiple model including single variables							AUC: 61.48(60.99–61.96) Sensitivity (FP10%): 24.73(24.04–25.44) AIC:117232				
Multiple model including combination of variables							AUC: 61.50(61.02–61.98) Sensitivity (FP10%): 24.72(24.05–25.39) AIC:117060				

AIC: Akaike information criteria, AF: Attributable fraction, AR: Attributable risk, TPF: True positive fraction, ORa: mutually adjusted Odds ratios. LR+: Positive Likelihood ratio; LR-: Negative Likelihood ratio, HBP: chronic hypertension; DM: Diabetes Mellitus

6. Dos últimos ejemplos

Veamos, para finalizar, una aplicación de los conceptos presentados hasta ahora sobre la capacidad discriminante de las pruebas diagnósticas utilizando como ejemplo los marcadores de enfermedad. El primero corresponde a las curvas ROC para conocer la capacidad discriminante de tres combinaciones de la expresión inmunohistoquímica de cuatro biomarcadores de un tumor (asignémosles los números del 1 al 3). La línea azul es la curva ROC para los biomarcadores 1 y 2 utilizados conjuntamente, la verde, la del 1, 2 y 3, y la morada, la del 1, 2, 3 y 4.



En principio, tampoco cabe duda de que deberíamos escoger la morada, que combina los marcadores 1, 2, 3 y 4, cuya ABC ronda 0,8 y es la mayor de las tres. En otras palabras, la combinación de los cuatro biomarcadores es la que tiene mayor capacidad para discriminar quién tiene y quién no tiene el tumor. Dicha ABC es mayor que la de la combinación de los biomarcadores 1, 2 y 3, cercana a 0,75 y que la de la combinación del 1 y 2, próxima a 0,70. No obstante, como carecemos de los IC95% de estas ABC (un error de subnotificación frecuente en publicación), y a juzgar por las imágenes, es posible que las tres ABC compartan bastantes valores de dichos IC95% y, por lo tanto, que la capacidad discriminante, cuando menos la de las combinaciones de dos y tres biomarcadores, no se diferencie sustancialmente.

Y el segundo ejemplo, también de preeclampsia, porque es un tema controvertido en la actualidad, y porque la presentación de las curvas ROC es mucho más ilustrativa que en el caso anterior. En la figura 2 del artículo [O’Gorman N, et al. Am J Obstet Gynecol 2016;214:103.e1-103.e12](#), disponible en la bibliografía, se presentan las curvas ROC de cuatro biomarcadores de eclampsia.

Lo que no presentan los autores del primer estudio, lo hacen los del segundo, lo cual facilita mucho la interpretación. Imaginemos que el factor de crecimiento placentario (FCP, curva ROC en rojo) es un nuevo marcador de preeclampsia y queremos decidir si debemos incorporarlo o no en nuestro servicio (deberíamos hacer lo mismo, por ejemplo, con un nuevo marcador tumoral). En la tabla aparecen los valores de las curvas ROC de cada uno de los cuatro marcadores y de sus combinaciones. Fijémonos solamente en las columnas de la izquierda, que corresponden a < 37 semanas de gestación. La información que sigue no se presenta en el artículo de los marcadores del melanoma.

A semejanza de cómo se presentaban los resultados en el estudio de los factores de riesgo cardiovascular, en la tabla podemos apreciar que la capacidad discriminante conjunta (ABC) de los factores maternos, la tensión arterial media, el índice de pulsatilidad de la arteria uterina, la proteína A plasmática asociada con el embarazo y el factor de crecimiento placentario sólo es 10,7 unidades más alta que la de los factores maternos aislados. De nuevo, es lógico y lo más razonable desde los puntos de vista de la biología y la epidemiología que con las combinaciones de signos, síntomas, pruebas diagnósticas complementarias y varios biomarcadores, no con ellos aisladamente, se alcance la mayor capacidad para discriminar resultados, respuestas, efectos.

Las ABC nos ofrecen información útil en este caso y en general para:

- 1) valorar si la magnitud de la capacidad discriminante incremental del signo, síntoma, de la prueba diagnóstica complementaria o del nuevo biomarcador, es decir respecto a los restantes conocidos considerados individualmente o en combinación, es alta o baja,
- 2) conociendo el precio y calculando el coste incremental de este marcador respecto a los restantes individualmente o en combinación, decidir si vale lo que cuesta,
- 3) calcular, si conocemos los precios, el coste marginal (cuánto cuesta aumentar en una unidad (1,0 el ABC) la capacidad discriminante de pre-eclampsia con el nuevo marcador aislado o combinado con otros respecto a los restantes considerados individualmente o combinados),
- 4) tener una aproximación de su impacto presupuestario, si disponemos de una estimación del total de personas a quienes se mediría el nuevo biomarcador, por ejemplo, en un año,
- 5) recordar que una baja capacidad discriminante responde en parte a nuestra actual falta de conocimiento completa o cabal sobre los mecanismos de causalidad entre las supuestas causas (o factores de riesgo) y los resultados o desenlaces o efectos. Con todo ello, dispondremos de información valiosa sobre el valor diagnóstico relativo o incremental de cualquier nuevo biomarcador, sin olvidar
- 6) **que la validez analítica sólo es un criterio de valoración de las pruebas diagnósticas y que dicha valoración, para ser completa, debe basarse en los tres criterios esenciales explicados (validez técnica, validez analítica y utilidad clínica), así como en los criterios de su evaluación económica.**

¿Cuántas decisiones de incorporación o no de nuevas pruebas diagnósticas, incluidos los biomarcadores, se toman sin disponer de toda esta información y qué consecuencias puede tener actuar así desde la perspectiva de la sociedad?

7. Recapitulación

Con estos ejercicios se ha intentado ilustrar:

- 1) La valiosa información adicional (el valor incremental de la información) que puede ofrecer el estudio de la capacidad discriminante de los supuestos factores de riesgo, así como su contribución al de la causalidad y la de las asociaciones entre ellos. Recordemos que, aunque los RR, OR o HR sean altos, su capacidad discriminante suele ser baja cuando consideramos factores aislados. No olvidemos que las RV^+ y $-$ de factores individuales suelen ser < 10 y $> 0,1$, lo que indica que estos factores no permiten confirmar ni descartar por sí solos la enfermedad. Esto es congruente con la realidad: ¿cuántos diagnósticos podemos confirmar o descartar con un solo signo o síntoma? Lo hacemos, primero, cuando tenemos información sobre varios signos y síntomas; en otras palabras, cuando las RV^+ y $-$ son > 10 y cercanas a $0,1$ respectivamente (suelen corresponder a las de combinaciones de signos y síntomas) y cuando el ABC es elevada ($> 0,8$).
- 2) El riesgo de errar que encierra extraer conclusiones y extrapolar más allá de lo que permiten los datos y la información disponible cuando con los estudios realizados sólo se tiene información parcial. No basta con estimar RR, OR o HR, *Association is not causation* ni necesariamente capacidad para discriminar individuos sanos de enfermos por el hecho de ser medidas promedio estimadas en grupos o poblaciones.
- 3) La necesidad de valorar los residuales, la bondad de ajuste (calibración) y la capacidad discriminante de los modelos multivariantes antes de extraer conclusiones sobre el significado real de la información que ofrecen y, menos aún, antes de tomar medidas o de iniciar intervenciones.
- 4) El hecho de que adoptar medidas y emprender acciones en las esferas de la clínica, las políticas de salud, los programas de prevención, y en macro, meso y microgestión basadas en información incompleta puede producir más daño que beneficios para la salud de poblaciones e individuos, como numerosos casos ponen de manifiesto; tampoco hay que caer en la parálisis por el análisis, porque la práctica cotidiana no nos deja.
- 5) **Que una baja capacidad discriminante pone de manifiesto dos hechos de suma importancia.** Primero, nuestra actual falta de conocimiento cabal sobre los mecanismos de causalidad entre las supuestas causas (o factores de riesgo) y los resultados o respuestas o desenlaces o efectos. Y, segundo, que dicho conocimiento es insuficiente porque con él (expresado por las variables con que intentamos predecir o discriminar) **no podemos captar o explicar todavía toda la heterogeneidad (la variabilidad) que existe y detectamos en los efectos (respuestas) a nivel individual ni la fracción de dicha heterogeneidad (variabilidad, varianza) entre individuos** que responde no sólo a factores propios de individuo sino a factores contextuales, del entorno. **(La varianza entre individuos no expresa incertidumbre, sino dicha heterogeneidad (biológica,**

incluso natural de los individuos) de respuestas a los distintos factores causales, la que no podemos explicar aun completamente.) Esto explica también, junto con la alta prevalencia de falsos positivos y falsos negativos en la población general, que la capacidad discriminante o predictiva a nivel individual (de respuesta, efecto, toxicidad, pronóstico) de muchos supuestos factores de riesgo y biomarcadores, aunque tengan asociaciones fuertes con las respuestas o efectos (OR, RR elevados), sea baja y, por consiguiente, su validez analítica sea todavía insuficiente.

En los últimos años, en medicina se están empezando a aplicar métodos de análisis e inteligencia artificial (agrupados bajo el concepto de *Machine Learning*) con grandes bases de datos (*big data*) y datos de la vida real (*real world data*), empleados en otras disciplinas desde hace años, que utilizan algoritmos muy complejos con los cuales se mejoran muchísimo las estimaciones de la capacidad discriminante. Entre ellos destacan los *classification trees* (C-TREE), los *classification and regression trees* (CART) y los *random forests*. (En la bibliografía se incluye una referencia que los describe de forma sencilla y muy intuitiva y los compara con los métodos de análisis estadísticos tradicionales).

8. Material didáctico complementario

Fayyad-Kazan J, et al. Circulating miR-150 and miR-342 in plasma are novel potential biomarkers for acute myeloid leukemia. J Translational Med. 2013;11:31. <http://www.translational-medicine.com/content/11/1/31>.

Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004; 159:882–90.

Garbe C, et al. Risk Factors for Developing Cutaneous Melanoma and Criteria for Identifying Persons at Risk: Multicenter Case-Control Study of the Central Malignant Melanoma Registry of the German Dermatological Society. Invest Dermatol. 1994;102:695-9.

Riera M, Amézaga R, Molina M, Campillo-Artero C, Sáez de Ibarra JI, Bonnín O, Ibáñez J. Mortalidad de las complicaciones postoperatorias (failure to rescue) tras cirugía cardiaca en un hospital universitario. Rev Cal Asist.2016;31:126-33.

Thomas KE, Hasbun R, Jekel J, Quagliarello VJ. The Diagnostic Accuracy of Kernig's Sign, Brudzinski's Sign, and Nuchal Rigidity in Adults with Suspected Meningitis. Clin Inf Dis. 2002;35:46-52.

O'Gorman N, et al. Am J Obstet Gynecol 2016;214:103.e1-103.e12.

9. Bibliografía

General

Nivel introductorio

- Gordis L. Epidemiology, Second Edition. Philadelphia, PA: WB Saunders Company; 2000.
- Mausner JS, Kramer S. Epidemiology: An introductory text. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. Foundations of epidemiology. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epiemiología clínica. Ciencia básica para la medicina. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. Epidemiology, Beyond the basics. Gaithersburg, MD: Aspen Publishers, Inc.; 2000.

Específica de este Capítulo

- Xunta de Galicia, Consellería de Sanidade e Servizos Sociais. Organización Panamericana de la Salud, Programa Especial de Análisis de Salud. Análisis Epidemiológico de Datos Tabulados (Epidat). Módulo de pruebas diagnósticas (incluye curvas ROC). <http://www.sergas.es/Saude-publica/Documents/1932/6-Ayuda%20Pruebas%20diagnosticas.pdf>
- Merlo J. Invited commentary. Multilevel analysis of individual heterogeneity—A fundamental critique of the current probabilistic risk factor epidemiology. Am J Epidemiol. 2014;180:208-12.
- Merlo J, Wagner P, Juarez S, Mulinari S, Hedblad B. (2013). The tyranny of the averages and the indiscriminate use of risk factors and population attributable fractions in Public Health: the case of coronary heart disease. Working paper version 2013-09-26. Unit for Social Epidemiology, Department of Clinical Sciences, Faculty of Medicine, Lund University, Sweden Malmö, Sweden.
- Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004; 159(9):882-90.
- Campillo-Artero C. The regulation of biomarkers. In: Rey-Lopez JC, Lumbreras B, Ponce-Lorenzo JJ, Campillo-Artero C, Maria Pastor-Valero. Precision medicine: Will technology be leveraged to improve population health? Amsterdam: Springer; 2019.

Machine Learning

- Breiman L. Statistical modelling: the two cultures. Stat Sci. 2001;16(3):199-231.

Capítulo 7.

Asociación, riesgo y causalidad

Índice

1. Introducción	118
2. <i>Association is not causation</i>	118
3. Interpretando asociaciones, riesgo y causalidad	120
4. Explicar y predecir	121
5. Causalidad inversa	122
6. Estudios que permiten estimar efectos causales en ausencia de aleatorización	122
7. Material didáctico complementario.....	131
8. Bibliografía.....	132

Objetivos específicos de este Capítulo

1. Mejorar el conocimiento sobre la causalidad.
2. Entender la diferencia entre asociación, riesgo y causalidad en epidemiología.
3. Prevenir de las consecuencias de las interpretaciones erróneas de estos tres conceptos cruciales en investigación, no sólo en medicina, y del abuso con que se hacen atribuciones de causalidad sin los fundamentos necesarios.

Nota importante

En este Cuaderno no se pueden reproducir figuras y tablas de artículos publicados protegidos, que, por su idoneidad, al tratarse de ejemplos prácticos claros, se utilizan para explicar algunos conceptos de los Capítulos 3, 6 Parte I y 6 Parte II y 7. Por este motivo, en cada caso se incluye la referencia bibliográfica donde se encuentran. La consulta exclusivamente de dichas tablas y figuras, guiada en las explicaciones, es imprescindible para entender esos conceptos; para ello no es necesario leer los artículos. La lista de las referencias correspondientes se incluye antes de la Bibliografía como **Material didáctico complementario**.

1. Introducción

En este último Capítulo, considerando todo lo explicado en los anteriores, se comentan tres temas de capital importancia en epidemiología y las relaciones entre ellos: asociación, riesgo y causalidad. Para entender este Capítulo es imprescindible haber estudiado los Capítulos precedentes.

Primero se insiste en dejar clara una máxima esencial en el estudio de la causalidad ya comentada: asociación no es causalidad (*Association is not causation*).

También se vuelve a revisar otro error grave y muy frecuente tratado en el Capítulo 6: considerar a ciertas variables de exposición como *factores de riesgo* cuando no lo son o no podemos saber si lo son con la información disponible.

Además, y no menos importante, se hace hincapié en algo explicado en otros Capítulos: los peligros que acarrea el abuso de establecer relaciones causales entre variables cuando se carece de suficientes pruebas científicas y justificaciones necesarias de índole metodológica, biológica o de otro tipo.

2. *Association is not causation*

Algo que debe quedarnos claro es que, en rigor, como señala MA Hernán en el artículo incluido en la bibliografía, las preguntas que nos formulamos en relación con la causalidad en epidemiología y en otras disciplinas son del tipo **¿cuál es el efecto causal medio de A en B estimado?** y que su respuesta son estimaciones de parámetros (por ej., coeficientes de regresión, como éstos se interpreten concretamente en cada tipo de modelo, OR, RR, HR etc.), junto con las medidas de incertidumbre asociadas con ellas (errores estándar, IC95%). **Las estimaciones siempre se deben acompañar de sus correspondientes medidas de incertidumbre. E, insistamos, sólo son estimaciones de efectos causales de A en B, no la confirmación de causalidad.** Realmente no formulamos la pregunta ¿es A la causa de B?, más de orden filosófico, y cuya respuesta es sí o no. De hecho, salvo muy raras excepciones en los millones de estudios publicados, en ninguna de estas disciplinas encontraremos estudios que se hayan realizado con mucho rigor que afirmen que A es la causa de B.

Simplificando un tema muy intrincado y controvertido, es extremadamente difícil poder afirmar sin duda que A es la causa de B. Lo que sí podemos hacer es *estimar efectos de naturaleza causal* de A en B y la incertidumbre de estas estimaciones, pero a sabiendas de que puede resultarnos científicamente muy difícil o imposible demostrar inequívocamente la causalidad. Algo estudiado es que del total de estudios que investigan causalidad, la fracción de los que aportan por consenso científico riguroso estimaciones “aceptables” de esos efectos es muy baja (muy bajo numerador y denominador altísimo).

Ello se explica por numerosas razones. Una muy frecuente, quizás la más frecuente, es que no se haya formulado adecuadamente la relación causal entre A y B que se pretende

estudiar: definición deficiente de la exposición o intervención A, del efecto causal B, incluidas las especificaciones relativas a quién, dónde, cuándo y cómo de esa supuesta relación causal. Si se parte de una formulación deficiente, los demás pasos del análisis arrastrarán esta deficiencia. Otras razones son diseños, métodos estadísticos, análisis e interpretación de resultados deficientes.

Supongamos ahora que con una formulación de la relación causal de la investigación y un diseño, conducción y análisis aceptables, se detecta la presencia de una asociación. Demos unos pasos atrás y recordemos lo siguiente: siempre que encontremos una asociación entre variables, o que no la encontremos, hemos de averiguar si es cierta o, por el contrario, si es el resultado del azar (error aleatorio), de sesgos o de factores de confusión (errores sistemáticos), tal como se ha expuesto en Capítulos anteriores.

Incluso disponiendo de una formulación suficientemente razonable y consensuada a nivel científico de las preguntas sobre relaciones causales, y habiendo descartado sesgos y factores de confusión *conocidos*, podemos aceptar la existencia una asociación potencialmente causal, pero debemos tener muy presente la *confusión residual* y, derivada de ello, la máxima anglosajona reiterada en epidemiología: *Association is not causation*, es decir, *Asociación no es causalidad*.

Confirmar la causalidad es una tarea muy ardua que exige muchos estudios y años de recopilación de información: con la incorporación de nuevos conocimientos, las preguntas sobre causalidad se reformulan (sólo a guisa de ejemplo, reparemos un momento en lo que está haciendo a este respecto el avance del conocimiento en genética) y, por lo tanto, nuestra concepción de dichas relaciones y su modelización con diseños y métodos de análisis disponibles. Los resultados obtenidos con la utilización de nuevos métodos de análisis (por ej., la inteligencia artificial) también nos hace replantear las preguntas sobre las relaciones causales entre una exposición y un efecto (un buen ejemplo y muy explícito lo ofrecen las investigaciones de los últimos años sobre la sepsis y alguno de los incluidos en el Capítulo 6 Parte II).

No pocos se han contentado con acotar la controversia sobre causalidad al cumplimiento o no de los criterios clásicos de causalidad, pero circunscribir el análisis sólo en torno a ellos, teniendo en cuenta que algunos de los cuales son controvertidos, es incurrir en una aproximación un tanto reduccionista. Tenerlos en cuenta y tener presente su cumplimiento es clave, pero, como puede colegirse de este capítulo, se debe ir más allá. Incluso reduciendo el análisis a estos criterios, se aprecia que la existencia de una asociación entre exposición y efecto y la fuerza de dicha asociación sólo es uno de ese conjunto de criterios. (Dichos criterios se encuentran en el Anexo 1.)

No escasean las intervenciones que se han llevado a cabo ni las recomendaciones de intervenir que se difunden de forma irresponsable, por desconocimiento, partiendo de falsos supuestos de causalidad, aunque sea de forma implícita, sin 1) haber descartado que la asociación encontrada sea espuria (fruto de errores aleatorios o sistemáticos: sesgos o factores de confusión), ni 2) haber comprobado que los supuestos factores de

riesgo lo son realmente, ni 3) haber confirmado el cumplimiento de los criterios de causalidad restantes, ni 4) considerar todo lo que en este cuaderno se trasmite sobre el conocimiento actual de la causalidad.

3. Interpretando asociaciones, riesgo y causalidad

En la tabla I del artículo Garbe C, et al. Risk Factors for Developing Cutaneous Melanoma and Criteria for Identifying persons at Risk: Multicenter Case-Control Study of the Central Malignant Melanoma Registry of the German Dermatological Society. J Invest Dermatol. 1994;102:695-9 se presentan los resultados de un modelo de regresión logística construido para averiguar qué variables pueden ser *factores de riesgo* de melanoma (CM). Repasemos algunos conceptos de estadística.

La medida de asociación entre cada variable y la presencia o no de melanoma utilizada en este estudio es el riesgo relativo. Se observa, por ejemplo, que a medida que aumentan el número de nevos comunes y atípicos, así como de léntigos actínicos, lo hace la fuerza de esta asociación, a juzgar por los valores de los RRs. Muchos concluirían *arriesgadamente* que son factores de riesgo de desarrollar un melanoma. En el modelo no se han incluido variables no asociadas con el desarrollo del tumor (los IC95% de sus RRs incluirían el valor 1,0) ni factores en principio supuestamente protectores (los RRs serían $< 1,0$ y todos los valores de sus IC95% serían $< 1,0$).

Recordemos que todos y cada uno de los RRs (o los coeficientes de regresión de los modelos como éstos se interpreten concretamente en cada tipo de modelo, y sus derivados: los OR o los HR de un modelo de supervivencia de Cox, si se hubiera construido en este caso) de cada variable están ajustados por todas las restantes variables incluidas en el modelo. Por ejemplo, los RRs de los nevos melanocíticos atípicos (1,6 y 6,1) están ajustados por las restantes cinco variables que están incluidas en este modelo.

Por otra parte, sus IC95% son prácticamente diferentes, sólo comparten un valor, 2,3, y por ello puede afirmarse que la fuerza de la asociación aumenta de magnitud, se multiplica prácticamente por cuatro, al pasar de 4 a 5 o más lesiones. Esto puede estar reflejando una relación dosis-respuesta, un criterio de causalidad como muestra el Anexo 1. Lo mismo se observa con los léntigos actínicos, aunque en este caso se solapan más valores de los IC95%.

Con esta información, desde los Capítulos 3, 4 y 5, y suponiendo que no han intervenido sesgos cuya magnitud sea suficiente para modificar los resultados (los efectos medios causales estimados), ya sabemos que no podemos descartar la *confusión residual* (*residual confounding*) y que, si hubiera, nos encontraríamos con estimadores sesgados (no reales) de los RR, la medida de la fuerza de asociación. Por consiguiente, y con suma cautela, que siempre debemos ejercer y hacer explícita, podemos suponer que hay asociaciones entre esas variables y la presencia de melanoma.

El siguiente paso, como se ha indicado en el Capítulo 6, deben analizarse las condiciones adicionales que deben cumplirse, además de la existencia de una asociación, para poder afirmar o no que estas variables son o no factores de riesgo de melanoma.

Aunque los resultados de los análisis precedentes permitan pensar que las variables consideradas son factores de riesgo de melanoma, nos falta mucha información para poder afirmarlo y mucha más sobre la causalidad de estas asociaciones; la han omitido los autores y no la han reclamado los editores.

4. Explicar y predecir

Al evaluar estos modelos siempre hemos de preguntarnos cuál es la finalidad de los autores. En general, un modelo puede haberse construido para conocer los factores que *explican* por qué hay pacientes que desarrollan un melanoma y otros no. No es infrecuente que estos carezcan de una formulación adecuada de la relación entre exposiciones y efectos en el sentido indicado más anteriormente sobre la causalidad.

Por otra parte, los investigadores podrían no estar interesados en explicar, por ejemplo, por qué unos pacientes con melanoma, cáncer de pulmón de células no pequeñas o hepatocarcinoma responden a un inhibidor de check point PD-1 como el pembrolizumab y otros no, sino en saber, únicamente, con qué variables se puede *predecir* razonablemente qué pacientes responderán y cuáles no.

A veces, en la práctica predecir y explicar son sinónimos: al seleccionar las variables predictivas de un efecto, de hecho, y de un modo u otro, hacemos razonamientos explicativos al relacionar las variables con los efectos. Pero en ocasiones no lo son exactamente, interesa más predecir que explicar, y podemos acabar teniendo modelos predictivos razonablemente aceptables de los efectos (con buena capacidad discriminante en conjunto) con los cuales no podamos explicar tan razonablemente dichos efectos (tratamos enfermedades con efectividad (predecimos efectos de medicamentos) cuya etiología y fisiopatología seguimos desconociendo).

Si dispusiéramos de todo el conocimiento necesario, es posible que las variables que *predicen* no sean exactamente las mismas que las que *explican*. Si queremos predecir si mañana saldrá el sol, basta con el encendido de las farolas de la calle el día antes para predecirlo. Pero con esta variable, aunque predecimos bien, no explicamos nada. Para explicar por qué mañana cabe esperar que salga el sol hemos de recurrir a la astronomía. Lo mismo ocurre en cualquier ejercicio en epidemiología: es preciso discernir claramente el objetivo de un estudio, porque las dos finalidades subrayadas aquí pueden no ser idénticas, como acostumbra serlo las variables y, a veces, los modelos que hemos de utilizar para un fin u otro.

En el Capítulo 6, Parte II, se repasan dos conceptos esenciales de los modelos construidos con fines predictivos, que también han de conocerse antes de extraer ninguna conclusión:

su bondad de ajuste o calibración y su capacidad discriminante. Como se ha indicado en él, puede ser un craso error afirmar exclusivamente sobre la base de los coeficientes de regresión, los valores de los ORs estimados por estos modelos, que una variable determinada es un factor de riesgo de la que se utilice como variable dependiente, en el ejemplo de más arriba la presencia de un melanoma, y puede ser aún más grave sin conocer su bondad de ajuste y su capacidad discriminante, y más aún si se interviene con los pacientes si en realidad no fuese un factor de riesgo.

No deben sorprender los numerosos estudios publicados en los cuales toda esta información no se notifica o se notifica parcialmente, y por esta razón debe dudarse de su validez interna y se ha de ser extremadamente cauto al extraer conclusiones de ellos y, sobre todo, al tomar decisiones o intervenir (probablemente con precipitación) respaldándonos en sus resultados.

5. Causalidad inversa

En los últimos años han aumentado el interés y los debates en torno al concepto de causalidad inversa. Veamos un ejemplo con la única finalidad de introducir este concepto en este cuaderno y por su importancia en epidemiología, dadas las consecuencias que puede tener recomendar intervenciones en clínica y en servicios de salud cuando se desconoce la presencia de causalidad inversa.

Desde hace años se considera que realizar ejercicio físico de forma mantenida en el tiempo es un factor protector de la enfermedad de Alzheimer, porque en diversos estudios las OR de la asociación entre actividad física y esta enfermedad son $< 1,0$. A falta de confirmación de su relación causal, en un estudio reciente en el cual se han utilizado métodos estadísticos más sofisticados que los habituales no se ha encontrado dicha asociación (los IC95% de los OR contienen el valor 1,0). Los autores interpretan este resultado del siguiente modo: los resultados de los estudios previos en que se encontró menor riesgo de enfermedad de Alzheimer en quienes realizan actividad física podrían ser atribuibles a causalidad inversa, es decir es la enfermedad en su fase preclínica la que produce una disminución de la actividad física en los que la están desarrollando.

6. Estudios que permiten estimar efectos causales en ausencia de aleatorización

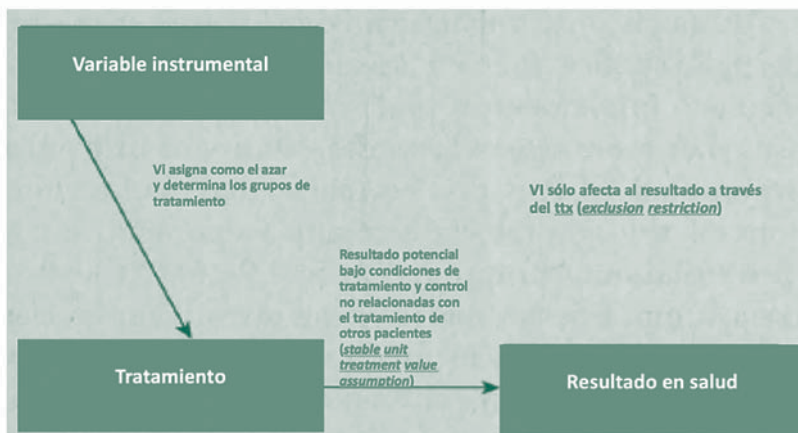
Es sumamente importante tener en cuenta que la clave al utilizar los estudios observacionales para estimar efectos causales es que deben concebirse y diseñarse intentando aproximarlos (remedar) en la medida de lo posible al ensayo clínico aleatorizado y controlado que *sería ideal* hacer pero no puede realizarse por numerosos motivos que conocemos y que nos conducen a los observacionales como única alternativa para estimar efectos causales. La validez de estos estudios observacionales (la de los efectos causales estimados con ellos) estriba en la medida en que con ellos se logre tener grupos de comparación lo más

parecidos posible a los que se consiguen con la asignación aleatoria, una tarea difícil, que no siempre se consigue. Esto último subraya la importancia decisiva de comprobar que se cumplen los requisitos específicos de cada estudio observacional en concreto. Aún y así, virtualmente nunca es posible descartar la existencia de confusión residual.

A continuación, se ofrece una descripción breve y solamente introductoria de tres diseños no experimentales (observacionales) y un método de emparejamiento con los cuales pueden estimarse efectos causales si se aplican adecuadamente, se cumplen todos sus requisitos y se minimizan las amenazas a su validez interna. En este artículo de acceso libre de la JAMA se encuentra una aproximación muy útil a este tema: <https://jamanetwork.com/journals/jama/fullarticle/205172>

6.1. Variables instrumentales

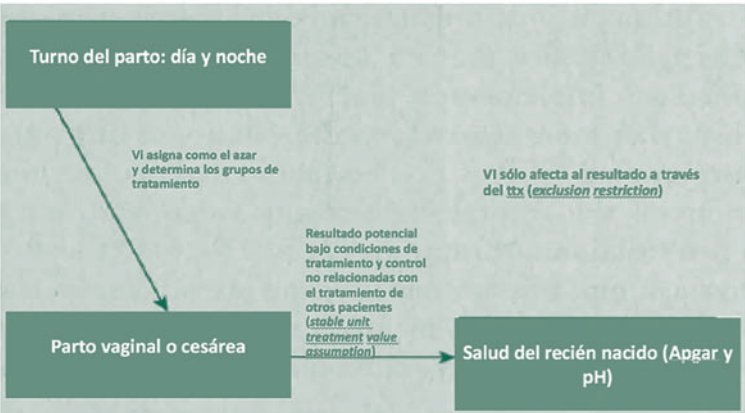
Una variable instrumental (VI) es aquella que, cuando se demuestra que se cumplen estrictamente varios requisitos, permite crear grupos asociados a intervenciones, tratamientos (o exposiciones) como lo hace la asignación aleatoria. En el siguiente esquema se ilustra cómo actúa la VI y se resumen dichos requisitos, que son, fundamentalmente, tres:



1. La VI asigna como lo haría el azar y determina los grupos de tratamiento en los cuales las distribuciones de las variables confusoras (*observables* y *no observables*) se distribuyen de forma muy similar, lo que permite *estimar* efectos causales de dichos tratamientos.
2. La VI sólo afecta al resultado a través del tratamiento, nunca directamente (criterio de restricción de la exclusión, *exclusion restriction*).
3. El resultado potencial que cabe esperar en cada participante que recibe el tratamiento no depende del tratamiento que recibe cada uno de los demás participantes, es independiente de él (supuesto del *valor estable de la unidad tratada*, *stable unit treatment value assumption*).

En este artículo se muestra un ejemplo de utilización de una VI: [Costa-Ramón AM, Rodríguez-González A, Serra-Burriel M, Campillo-Artero C. It's about time: Cesarean sections and neonatal care. J Health Econ. 2018;59:46-59](#). En él se quiere averiguar si las cesáreas causan daño al recién nacido comparadas con el parto vaginal eutócico, no instrumental. Como puede colegirse de inmediato, la pregunta que late es si un efecto causal de la cesárea es el daño al recién nacido. Si la VI “funciona”, como se demuestra en este estudio al comprobar el cumplimiento de los tres requisitos, el efecto estimado de las cesáreas en el recién nacido utilizando como variables de resultado o efectos la prueba de Apgar y el pH intraparto, puede considerarse de naturaleza causal, como se haría en un ensayo clínico, que, como cabe esperar, no puede realizarse para estimarlo, algo ya comentado en Capítulos anteriores. El uso de VI constituye, por lo tanto, un estudio observacional (los datos están registrados *ex ante* y no se interviene).

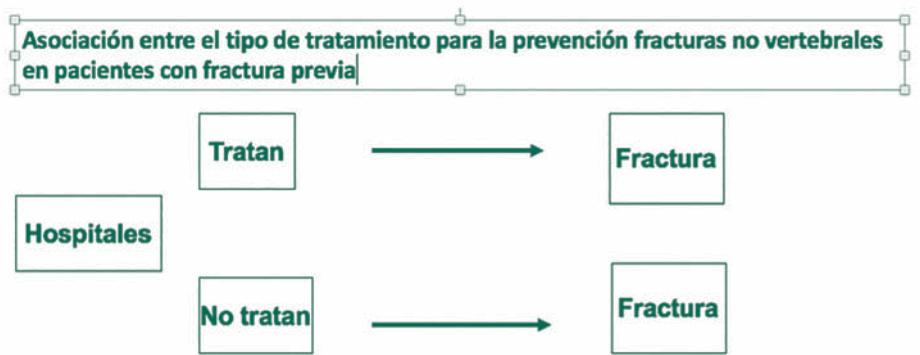
El siguiente esquema ilustra el uso de la VI que se escogió: el tramo horario de realización de cesáreas, por la noche (*night shift*), entre las 11 de la noche y las 3 de la mañana (cuando se observa un pico diario de cesáreas en muchos hospitales, incluyendo y excluyendo las inducciones), y el resto del día.



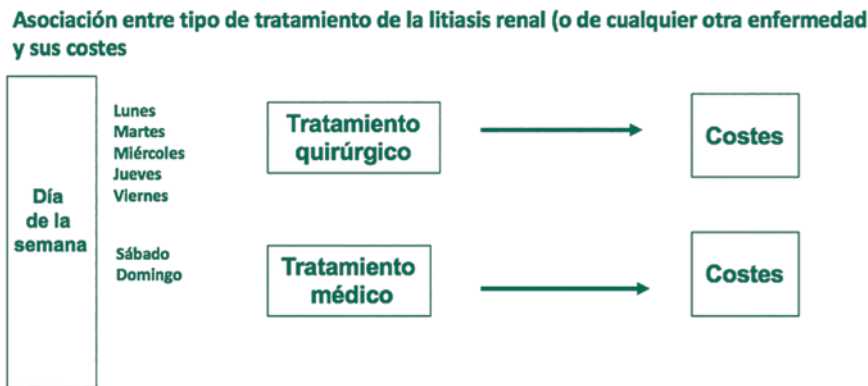
La tabla 1 del artículo indica que la distribución de las variables potencialmente confusoras del efecto escogido (salud del recién nacido) en el turno de noche y en el resto del día son prácticamente iguales, salvo el estado civil (que no es confusora) y la inducción (más frecuente de día, pero que, al repetir los análisis con y sin las inducciones, se obtienen los mismos resultados, lo que excluye que sea confusora). Esas distribuciones lo son también al conformar los dos grupos que se quieren comparar, los partos vaginales y las cesáreas. Además, mediante varios análisis y ajustes se comprueba que se cumplen los otros dos requisitos (la restricción de la exclusión y el valor estable de la unidad tratada), con lo cual se puede proceder a estimar los efectos del tipo de parto en la prueba de Apgar y el Ph intraparto, como se haría tras una asignación aleatoria en un ensayo clínico.

Los resultados indican, como muestran varias tablas, que, si bien hay un efecto estadísticamente significativo (se incluyeron 6163 partos), su significación clínica es irrelevante (descenso medio en las cesáreas respecto de los partos vaginales de la prueba de Apgar y del pH, ingresos en la UCIN y mortalidad).

Otras VI que se utilizan, a modo de ejemplo, son las siguientes:



En este caso la VI es el hospital: unos hospitales tratan con X y otros, no (indican otro tratamiento), y podríamos hacer lo mismo con tratamientos quirúrgicos. Procedemos del mismo que en el caso anterior: si las distribuciones de las variables potencialmente confusoras son muy similares en los dos grupos (tratados con un medicamento X y no tratados con él sino con otro dependiendo del hospital) y se comprueba el cumplimiento de los requisitos indicados, puede estimarse el efecto causal (efectividad y seguridad) del tratamiento X respecto al otro (la distribución de las variables observables y no observables es muy similar), si bien siempre hay que admitir la posibilidad de confusión residual.



Y en este ejemplo la VI es el día de la semana: en uno o en varios hospitales tratan quirúrgicamente la litiasis renal de lunes a viernes e indican tratamiento médico los fines de semana. Siempre procederíamos como en los casos anteriores.

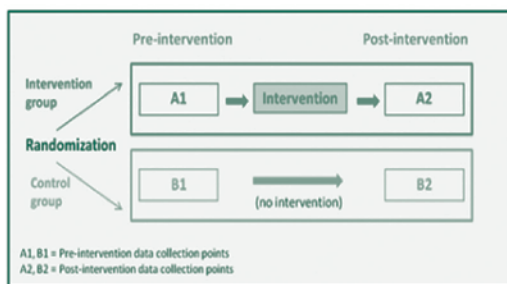
No es fácil encontrar VI adecuadas, en esto radica su principal problema, pero cuando se encuentran, sirven para el propósito para el cual se utilizan y muchos investigadores optan por ellas como el tipo de estudio observacional que mejor substituye a la asignación aleatoria. En ese artículo se encuentran varios ejemplos de VI: [Faden Garabedian L, Chu P, Toh S, Zaslavsky AM, Soumerai SB. Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research. Ann Intern Med. 2014;161:131-8.](#)

6.2. Estudio cuasi-experimental pre-post con grupo control (diferencia en diferencias)

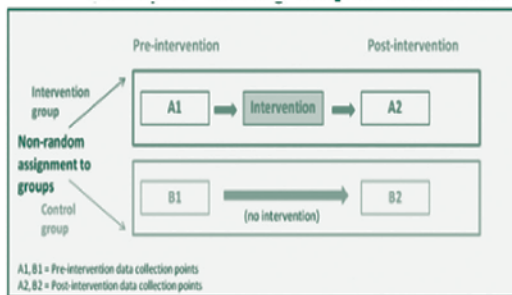
Este estudio forma parte de un conjunto de diseños en los cuales la formación de los grupos se realiza **sin asignación aleatoria**. Por este motivo, se llaman **cuasi-experimentales**. Veamos en qué consiste con los siguientes esquemas.

Sabemos de Capítulos anteriores que en los diseños experimentales (como los ensayos clínicos) los grupos experimental y control se forman por asignación aleatoria (A1 y B1, esquema de la izquierda). Se realizan mediciones basales (pre, antes de la intervención) en ambos grupos (por ej., administrar un medicamento, implantar un sistema de copago creciente en diferentes grupos de personas aseguradas (formados por asignación aleatoria) para ver si influyen en la demanda de medicamentos, como se hizo en el experimento de la RAND Corporation), se realiza la intervención y, cuando esta finaliza, se hacen mediciones (post, A2, B2) en los dos grupos y se comparan para ver si hay diferencias entre ellos, lo que constituye el efecto de la intervención (la estimación de su efecto causal).

Diseño experimental clásico



Diseño cuasi-experimental



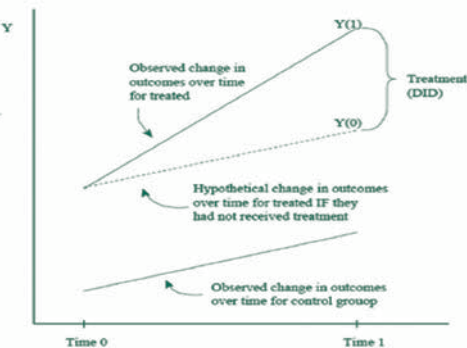
<https://www.k4health.org/toolkits/measuring-success/types-evaluation-designs>

En el cuasi-experimental se selecciona el grupo en que se va a intervenir y se busca un grupo control (un grupo lo más parecido al de intervención en la distribución de las variables asociadas o potencialmente asociadas con los resultados y se consideran de confusión), se hacen igualmente mediciones basales (pre, A1 y B1), se interviene en el grupo de intervención, y cuando esta acaba, se hacen de nuevo mediciones (post, A2, B2), y se comparan, igual que en los experimentales, para ver si hay diferencias entre ellos, lo que constituye la estimación del efecto de la intervención.

Al hacer estas comparaciones deben controlarse siempre los factores de confusión (y aceptar siempre la posibilidad de *confusión residual*). Asimismo, puede ser necesario controlar cambios en el propio grupo de intervención o el efecto de factores exógenos (al margen de la intervención) concomitantes que puedan haber influido en los resultados durante el periodo de intervención (por ejemplo, cambios importantes en el personal del servicio clínico, un cambio sustancial en su modelo de gestión, la introducción de una nueva tecnología o la existencia de **cointervenciones**, por ejemplo, si se está evaluando pre-post la efectividad de un programa de reducción de la infección nosocomial, la implantación concomitante de un nuevo protocolo de antibioticoterapia y del lavado de manos con solución hidroalcohólica de forma generalizada en el hospital).

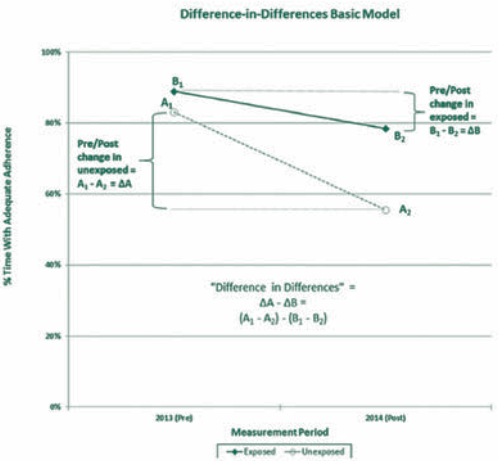
Por todo ello, es necesario un **grupo control**, que, como es muy parecido al de intervención tal como lo escogemos y tras los ajustes señalados, nos informa sobre el *contrafactual*, lo que le habría ocurrido al de intervención si en este último no se hubiera intervenido. Esto cual permite separar los efectos observados atribuibles a la intervención de los que son consecuencia de esos de confusión y de los exógenos, cambios importantes que hayan ocurrido durante la intervención en la distribución de las variables del grupo de intervención, que estén asociados con los resultados (tanto más probables cuanto más largo sea el periodo de la intervención y viceversa).

A continuación, se compararan los resultados post (la variable o las variables de resultado escogidas: proporciones, tasas, RR, OR, HR, etc.) de ambos grupos (intervención y control), para averiguar si hay diferencias entre ellas (el efecto de la intervención), que, con este diseño, al ser los dos grupos muy parecidos respecto a la distribución de aquellas variables mencionadas, habiendo ajustado las variables post si ha sido necesario, controlando por los factores de confusión y cumpliéndose varios supuestos específicos que tienen estos diseños, pueden atribuirse también (efectos causales estimados) a la intervención, del mismo modo que se atribuye a un medicamento el efecto observado en el cambio post en un ensayo clínico. Veámoslo en los dos esquemas de este diseño, que también se llama de diferencia en diferencias (*difference-in-differences*): https://www.lexjansen.com/wuss/2016/49_Final_Paper_PDF.pdf:



En el de la izquierda se aprecia el cambio pre (Time 0) — post (Time 1) observado en los resultados en el grupo intervenido Y (1) y el esperado en él si no se hubiese intervenido Y(0) (la línea de puntos, cuya pendiente es paralela a la del grupo control, representado por la línea de abajo).

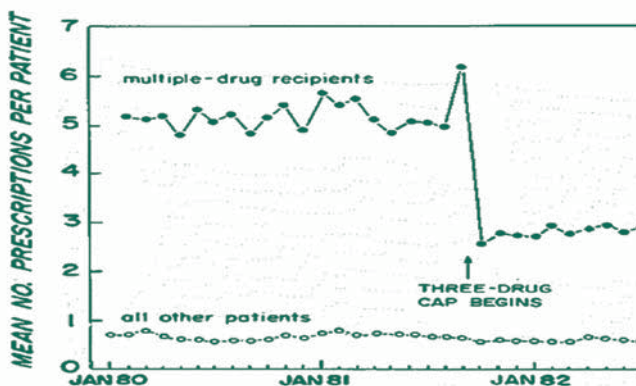
En este otro ejemplo, se trata de una intervención dirigida a frenar el descenso de la adherencia a un tratamiento con el paso del tiempo. El efecto de la intervención es $B_1 - B_2$. En cambio, en el grupo control (no intervenido) se ha ido produciendo un descenso de la adherencia de mayor magnitud que en el intervenido (la pendiente descendente es más pronunciada que la del grupo intervenido). Dicha magnitud es la diferencia $A_1 - A_2$. El efecto estimado de la intervención es la diferencia entre ambas diferencias $= (A_1 - A_2) - (B_1 - B_2)$.



La descripción de los modelos multivariantes que se construyen para estimar los efectos ajustados no es objeto de este cuaderno introductorio.

6.3. Series temporales interrumpidas

La figura que aparece a continuación muestra el cambio en una serie temporal de la media del número de prescripciones por paciente y por mes en el servicio de un hospital. Como esta media era muy alta, se han implantado varias medidas para reducirla y, tras un ascenso muy breve por las expectativas generadas al anunciarse la intervención, se observa un descenso brusco, muy pronunciado y duradero.



<https://rd.springer.com/content/pdf/10.1007%2Fs11606-016-3841-9.pdf>. Acceso abierto.

Para construir estas series temporales se ha de disponer de varias mediciones en el segmento pre y varias en el post. En este ejemplo, se aprecian dos segmentos, uno antes (pre) y otro después (post) de la intervención: de enero del 80 a julio del 81 y de esta última fecha a julio del 82. En ellas se mide el **nivel** (los valores o magnitudes de la variable que se estudie en varios intervalos de tiempo y en los dos segmentos, pre y post) y la **tendencia** o pendiente (el cambio de dicho nivel tras la intervención y a lo largo de cada segmento). En las series temporales interrumpidas, el nivel y la pendiente de la variable estudiada (en este ejemplo, la media del número de prescripciones por paciente y mes) en el segmento preintervención se comparan con el nivel y la pendiente del segmento postintervención.

Con modelos de regresión llamada segmentada se comparan los cambios del nivel y de la tendencia de la variable de resultado del segmento pre y post intervención. El cambio pre-post intervención del nivel de la variable constituye el efecto de la intervención. El cambio en la tendencia puede ser un aumento o un descenso de la pendiente tras la intervención o en un segmento.

Incluso sin tener un grupo control, (en este estudio pueden ser todos los pacientes restantes, *all other patients*), aunque con muchísimas prevenciones, este análisis puede sortear el efecto de algunas amenazas a la validez interna, porque se realizan varias mediciones

del efecto antes y varias después de la intervención, a diferencia de diseños menos protegidos de estas amenazas, como los transversales (que comparan el resultado en una sola medición simultánea en cada grupo) o los diseños cuasi-experimentales pre-post (con una sola medición pre y una sola post).

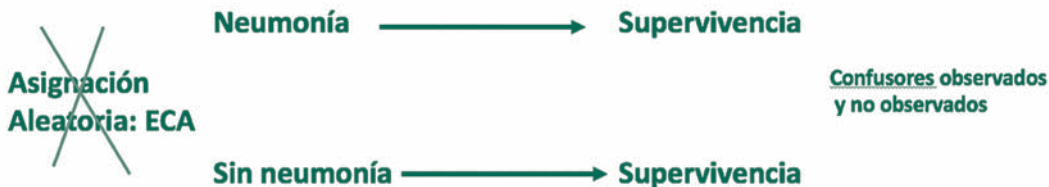
Los factores de confusión también deben controlarse incluyéndolos en los modelos al igual que las variables asociadas con los resultados que hayan cambiado en el momento de realizar la intervención, como intervenciones concomitantes (cointervenciones), cambios estacionales (periódicos) de la variable de resultado que hayan ocurrido, precisamente, en el momento de la intervención, cambios en la composición del grupo de intervención (como algunos enumerados en el diseño anterior) o en la forma de medir el resultado durante la intervención. Con la inclusión de un grupo control adecuado (hay varias opciones para hacerlo), a semejanza de los estudios de diferencia en diferencias, realizando los ajustes necesarios y cumpliendo con los supuestos de estos estudios (por ejemplo, la ausencia de correlación entre mediciones o que la tendencia y sus cambios son lineales) y utilizando métodos estadísticos adecuados cuando se incumplen los supuestos, se puede minimizar en buena medida el efecto de estas variables confusoras y de otras amenazas a su validez interna y estimar efectos causales de la intervención, aunque siempre, insistamos en ello, a sabiendas 1) de la presencia potencial de sesgos y de confusión residual, máxime porque la asignación de los participantes a los grupos de intervención y control no ha sido aleatoria, con lo que ello ya sabemos que supone y 2) de que estamos estimando efectos causales, no confirmando causalidad.

Hay numerosos ejemplos publicados del uso de estos diseños, como para estimar el efecto de los copagos en el consumo de medicamentos afectados por él, el de medidas de mejora de los cuidados de pacientes con infarto de miocardio o con ictus en las ambulancias de traslados a los hospitales, etc.

6.4. Propensity score

El propensity score (PE) es un método de emparejamiento de los participantes de los grupos en estudio. Expliquémoslo con un ejemplo con datos reales de este estudio: [Ibañez J, Riera M, Amezaga R, Herrero J, Colomar C, Campillo-Artero C, Saez de Ibarra JI, Bonnin O. Long-Term Mortality After Pneumonia in Cardiac Surgery Patients: A Propensity-Matched Analysis. J Intensive Care Med. 2014;31\(1\):34-40.](#)

El objetivo de este artículo es averiguar si hay diferencias en la supervivencia entre los pacientes que desarrollan una neumonía tras cirugía cardíaca y los que no desarrollan neumonía, es decir si un *efecto causal* de la neumonía postcirugía cardíaca es la reducción de la supervivencia. La opción ideal es hacer un ensayo clínico aleatorizado y controlado, pero, como ocurre en la mayoría de los casos, no es posible realizarlo. Además, podemos tener suficientes datos registrados para hacer algún estudio observacional con el cual, con las prevenciones que ya sabemos que hemos de tener en ausencia de aleatorización, podamos *estimar* ese efecto causal. El diseño es este:



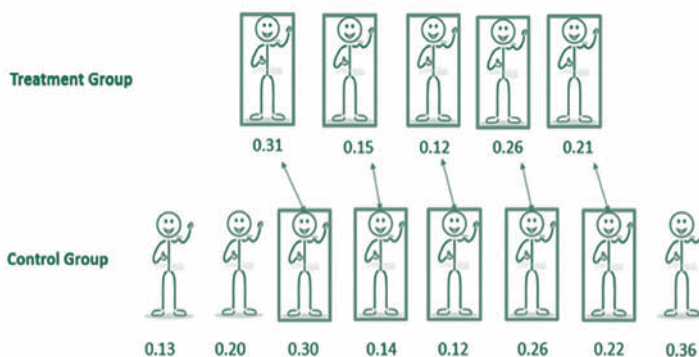
En ausencia de asignación aleatoria, ¿cómo sabremos si la diferencia en supervivencia se debe a la neumonía (hemos observado que sobreviven menos los que tienen neumonía) o a que la distribución (frecuencia) de pacientes mayores, con más complicaciones, más gravedad, IRC, cirugía cardíaca previa o anemia es diferente entre ambos grupos (más alta en los que padecen neumonía)? Son factores de confusión conocidos (*observables*).

Pues comprobamos si la distribución de estas posibles variables confusoras entre los que tienen y no tienen neumonía es o no similar. Si es diferente, pueden ser confusoras. Si lo son, ya sabemos que debemos ajustar las asociaciones por ellos (análisis estratificado o ajuste en modelos multivariantes adecuados) y que también podemos **emparejar** a los pacientes de cada grupo en esas variables, algo muy difícil de hacer “a mano” (muchos pacientes y a menudo unas cantas variables). La opción es la siguiente (en este libro importa el qué se hace, no el cómo desde el punto de vista cuantitativo).

Con una regresión logística se calcula el *propensity score*, como variable dependiente, que **se define como la probabilidad de desarrollar neumonía y, en general, como la probabilidad de recibir un tratamiento, en función de las variables de confusión seleccionadas.**

Acto seguido, se aparean pacientes con y sin neumonía con PS igual o muy cercano, para que la distribución de esas variables en ambos grupos sea lo más parecida posible (con ello eliminamos su potencial *confounding*). Con esto conseguimos que lo que los diferencie no sean esas variables sino solamente tener o no neumonía (más lo que desconocemos, la confusión residual), algo que, si funciona bien, lo que no siempre ocurre, **se aproxima a lo que se conseguiría con la asignación aleatoria.**

A continuación, se comprueba si con el PS hemos conseguido tener dos grupos (con y sin neumonía) con distribuciones muy similares de esas variables confusoras. Y, aunque no sea el caso en este estudio que sirve de ejemplo, deberíamos hacer lo mismo con otras de ajuste que usemos (por ej., una variable *dummy* para hospitales si el estudio fuese multicéntrico para controlar el efecto de hospital). Es lo que se observa en la Tabla 2 del artículo citado. Como se muestra, el PS ha funcionado: las distribuciones son muy parecidas. Hemos conseguido algo muy cercano a lo que logramos con la asignación aleatoria. Podemos proceder a *estimar* el efecto causal, (ojo, de nuevo, a estimar dicho efecto, no a confirmar causalidad entre neumonía y supervivencia, porque no podemos descartar la confusión residual).



Seguidamente, se construye un modelo adecuado (en este caso de supervivencia de Cox) con la supervivencia como variable dependiente, tener o no tener neumonía, el PS y otras variables de ajuste si fuese necesario, como se acaba de señalar, como independientes y el término error. Al introducir el PS como una variable independiente más del modelo, la asociación entre la neumonía y la supervivencia está ajustada por los factores de confusión que contiene, que captura el PS, como si hubiésemos introducido cada uno de ellos como variables independientes, y por todas demás variables independientes que queden en el modelo (se muestra en la figura 1 del artículo). Como ya sabemos, el efecto causal estimado con el modelo (que mide la beta de la variable dicotómica tener/no tener neumonía y la supervivencia, en este caso la *hazard ratio*), está exento de los sesgos y confusores observados que hemos considerado (conocidos y medidos, *observables*), *no de los no observados* (confusión residual).

Hay que tener muy presente que los PS no siempre funcionan “adecuadamente” y que para estudios observacionales hay investigadores que recomiendan utilizar en su lugar variables instrumentales, cuando se pueda identificar una que cumpla con todos sus requisitos, tarea a menudo difícil, porque los efectos causales estimados con ellas son más cercanos aún a los estimados con estudios aleatorizados porque con ellas se consigue reducir aún más la confusión residual que con otros métodos.

7. Material didáctico complementario

Garbe C, et al. Risk Factors for Developing Cutaneous Melanoma and Criteria for Identifying persons at Risk: Multicenter Case-Control Study of the Central Malignant Melanoma Registry of the German Dermatological Society. *J Invest Dermatol.* 1994;102:695-9.

Ibañez J, Riera M, Amezága R, Herrero J, Colomar C, Campillo-Artero C, Saez de Ibarra JJ, Bonnin O. Long-Term Mortality After Pneumonia in Cardiac Surgery Patients: A Propensity-Matched Analysis. *J Intensive Care Med.* 2014;31(1):34-40.

Costa-Ramón AM, Rodríguez-González A, Serra-Burriel M, Campillo-Artero C. It's about time: Cesarean sections and neonatal care. *J Health Econ.* 2018;59:46-59.

Faden Garabedian L, Chu P, Toh S, Zaslavsky AM, Soumerai SB. Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research. *Ann Intern Med*. 2014;161:131-8.

8. Bibliografía

General

Nivel introductorio

- Gordis L. *Epidemiology*, Second Edition. Philadelphia, PA: WB Saunders Company; 2000.
- Mausner JS, Kramer S. *Epidemiology: An introductory text*. Philadelphia, PA: WB Saunders Company; 1985.
- Lilienfeld AM. *Foundations of epidemiology*. New York: Oxford University Press; 1976.
- Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Epidemiología clínica. Ciencia básica para la medicina*. 2ª ed. Madrid: Editorial Médica Panamericana; 1994.
- Guías para revisar y publicar diferentes tipos de estudios: equator-network.org

Nivel intermedio

- Szklo M, Nieto J. *Epidemiology, Beyond the basics*. Gaithersburg, MD: Aspen Publishers, Inc. 2000.

Anexo 1. Criterios de causalidad clásicos

Fuerza de la asociación. Cuanto mayor sea la fuerza de la asociación estimada con RR, OR o HR, y si dichas estimaciones no son espurias, más probable es que la asociación sea causal.

Especificidad. Hace referencia a la “exclusividad” de la relación entre la supuesta causa y el efecto, al hecho de que la exposición a un factor concreto se asocia exclusivamente con una enfermedad. Un claro ejemplo son las enfermedades infecciosas en las cuales la causa es inequívocamente una bacteria, un hongo o un virus. Si la asociación es específica, es más posible que la asociación sea causal, aunque su ausencia no la excluye, porque algunos efectos (enfermedades, por ejemplo) responden a varias causas o a interacciones complejas entre distintas causas (multicausalidad) y porque la exposición a un factor concreto puede asociarse con enfermedades diferentes. Por estas razones, la especificidad es un criterio débil de causalidad y su ausencia no la excluye salvo en algunos casos.

Consistencia o replicabilidad de los resultados. El hecho de haber observado dicha asociación en otros estudios, con grupos diferentes y en situaciones y circunstancias distintas va a favor de una relación causal.

Plausibilidad. Poder explicar desde la biología, la fisiopatología u otra disciplina los mecanismos de la asociación causal es otro criterio de causalidad, si bien no siempre podemos explicar estas asociaciones.

Analogía. Alude a la similitud con otras asociaciones causales conocidas. Si un determinado factor es la causa de cierta enfermedad o efecto, es posible que factores con características o mecanismos de acción similares a él puedan producir el mismo o similares efectos.

Coherencia. Con este criterio se intenta averiguar si la asociación causal supuesta es coherente con el conocimiento disponible sobre la enfermedad o el efecto y con lo que se sabe respecto a su historia natural y demás factores relacionados con su desarrollo.

Pruebas experimentales. La existencia de pruebas experimentales es esencial para demostrar la relación causal, aunque a menudo se carece de ellas.

Gradiente biológico o relación dosis-respuesta. Si a medida que aumenta la intensidad de la exposición (mayor exposición a radiaciones ionizantes), mayor es la fuerza de la asociación (incidencia de leucemia y mayores RR, OR o HR), más probable es que la asociación sea causal. Sin embargo, su ausencia no excluye causalidad, porque en ciertos casos la asociación (la relación dosis-respuesta) no es lineal y sólo se detecta por encima de un determinado dintel de intensidad de exposición, como todo o nada.

Relación temporal. La causa debe preceder al efecto. No siempre es fácil demostrarlo. Por ejemplo, en cierto momento se sospechó que la asociación entre cimetidina y cáncer gástrico podría ser causal, porque en estudios de casos y controles el riesgo de cáncer gástrico en los expuestos a este medicamento era mayor que en los no expuestos a juzgar por los OR. La respuesta a la siguiente pregunta da la clave de la importancia de este criterio: ¿la ingesta de cimetidina precede al desarrollo de cáncer gástrico o los pacientes con este cáncer (ya sintomático) toman cimetidina para paliar la acidez que este tumor produce?

Referencias del material didáctico complementario

Esta es la relación de las referencias bibliográficas necesarias para consultar algunas de sus tablas y figuras y poder seguir algunas explicaciones de las Unidades 3, 6 Parte I, 6 Parte II y 7. Asimismo, se encuentran en la sección Material didáctico complementario de cada una de estas Unidades, antes de la Bibliografía.

UNIDAD 3. Ensayos clínicos

Stupp R, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* 2009;10:459–66.

Harrison CN, et al., Ruxolitinib versus Best Available Therapy for ET intolerant or resistant to Hydroxycarbamide in a Randomized trial. *Blood.* 2017;26;30:1889-97.

Fojo T, Mailankody S, Andrew Lo A. Unintended Consequences of Expensive Cancer Therapeutics—The Pursuit of Marginal Indications and a Me-Too Mentality That Stifles Innovation and CreativityThe John Conley Lecture. *JAMA Otolaryngology–Head Neck Surg.* 2014;140:1225-36.

Freidlin B, et al. Two-by-Two Factorial Cancer Treatment Trials: Is Sufficient Attention Being Paid to Possible Interactions? *JNCI.* 2017;109. doi: 10.1093/jnci/djx146.

Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. *NEJM.*2017;377:62-70.

Park JW, et al. Adaptive Randomization of Neratinib in Early Breast Cancer *N Engl J Med* 2016;375:11-22.

Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being missed? *Ann Intern Med.* 1996;125:605-13.

Montori VM, et al Validity of composite end points in clinical trials. *BMJ.* 2005;330:594-6.
Robert C, et al. Pembrolizumab versus Ipilimumab in Advanced Melanoma. *N Engl J Med.* 2015;372:2521-32.

Sun X, Briel M, Walter SD, Guyat GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ.* 2010;340:c117.

UNIDAD 6. Pruebas diagnósticas. Parte II

Fayyad-Kazan J, et al. Circulating miR-150 and miR-342 in plasma are novel potential biomarkers for acute myeloid leukemia. *J Translational Med.* 2013;11:31. <http://www.translational-medicine.com/content/11/1/31>

Pepe MS, Janes H, Longton G, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol.* 2004;159:882–90.

Garbe C, et al. Risk Factors for Developing Cutaneous Melanoma and Criteria for Identifying Persons at Risk: Multicenter Case-Control Study of the Central Malignant Melanoma Registry of the German Dermatological Society. *Invest Dermatol.* 1994;102:695-9.

Riera M, Amézaga R, Molina M, Campillo-Artero C, Sáez de Ibarra JI, Bonnín O, Ibáñez J. Mortalidad de las complicaciones postoperatorias (failure to rescue) tras cirugía cardíaca en un hospital universitario. *Rev Cal Asist.* 2016;31:126-33.

Thomas KE, Hasbun R, Jekel J, Quagliarello VJ. The Diagnostic Accuracy of Kernig's Sign, Brudzinski's Sign, and Nuchal Rigidity in Adults with Suspected Meningitis. *Clin Inf Dis.* 2002;35:46-52.

O'Gorman N, et al. *Am J Obstet Gynecol* 2016;214:103.e1-103.e12.

UNIDAD 7. Asociación, riesgo y causalidad

Garbe C, et al. Risk Factors for Developing Cutaneous Melanoma and Criteria for Identifying persons at Risk: Multicenter Case-Control Study of the Central Malignant Melanoma Registry of the German Dermatological Society. *J Invest Dermatol.* 1994;102:695-9.

Ibáñez J, Riera M, Amezaga R, Herrero J, Colomar C, Campillo-Artero C, Saez de Ibarra JI, Bonnin O. Long-Term Mortality After Pneumonia in Cardiac Surgery Patients: A Propensity-Matched Analysis. *J Intensive Care Med.* 2014;31(1):34-40.

Costa-Ramón AM, Rodríguez-González A, Serra-Burriel M, Campillo-Artero C. It's about time: Cesarean sections and neonatal care. *J Health Econ.* 2018;59:46-59.

Faden Garabedian L, Chu P, Toh S, Zaslavsky AM, Soumerai SB. Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research. *Ann Intern Med.* 2014;161:131-8.



RRI17035



9 788409 131907

