Cómo leer un artículo científico

LAS BASES DE LA MEDICINA BASADA EN LA EVIDENCIA

QUINTA EDICIÓN

TRISHA GREENHALGH

ELSEVIER

Cómo leer un artículo científico

Las bases de la medicina basada en la evidencia

QUINTA EDICIÓN

Trisha Greenhalgh
Professor of Primary Health Care
Barts and the London School of Medicine and Dentistry
Blizard Institute
London, UK

ELSEVIER
Ámsterdam Barcelona Beijing Boston Filadelfia Londres Madrid
México Milán Múnich Orlando París Roma Sídney Tokio Toronto

ELSEVIER

Edición en español de la quinta edición de la obra original en inglés
How to Read a Paper. The Basics of Evidence-Based Medicine

Revisión científica:
Pablo Alonso Coello
Investigador asociado
Centro Cochrane Iberoamericano
Instituto de Investigación Biomédica Sant Pau (IIB-Sant Pau), Barcelona

ISBN edición original: 978-1-118-80096-6
ISBN edición española (versión impresa): 978-84-9022-945-3
ISBN edición española (versión electrónica): 978-84-9022-973-6
Depósito legal (versión impresa): B. 20.167 - 2015
Depósito legal (versión electrónica): B. 20.168 - 2015

Servicios editoriales: DRK Edición

Impreso en Polonia

Advertencia

La medicina es un área en constante evolución. Aunque deben seguirse unas precauciones de seguridad estándar, a medida que aumenten nuestros conocimientos gracias a la investigación básica y clínica habrá que introducir cambios en los tratamientos y en los fármacos. En consecuencia, se recomienda a los lectores que analicen los últimos datos aportados por los fabricantes sobre cada fármaco para comprobar la dosis recomendada, la vía y duración de la administración y las contraindicaciones. Es responsabilidad ineludible del médico determinar la dosis y el tratamiento más indicado para cada paciente en función de su experiencia y del conocimiento de cada caso concreto. Ni los editores ni los directores asumen responsabilidad alguna por los daños que pudieran generarse a personas o propiedades como consecuencia del contenido de esta obra.

El editor

En noviembre de 1995, mi amiga Ruth Holland, editora de reseñas de libros en el *British Medical Journal,* me sugirió que escribiera un libro para desmitificar un tema tan importante, pero a menudo inaccesible, como el de la medicina basada en la evidencia. Ruth aportó valiosos comentarios sobre el borrador inicial del manuscrito, pero falleció trágicamente en un accidente ferroviario el 8 de agosto de 1996. Este libro está dedicado a su memoria.

Prólogo del profesor Sir David Weatherall a la primera edición

No es sorprendente que la gran cobertura publicitaria que se ha dado a lo que ahora se denomina *medicina basada en la evidencia* haya sido recibida con reacciones contrapuestas por parte de quienes están involucrados en la provisión de asistencia al paciente. La mayoría de los médicos parecen sentirse ligeramente heridos por el concepto, lo que sugiere que hasta hace poco toda la práctica médica ha sido lo que Lewis Thomas ha descrito como una especie frívola e irresponsable de experimentación humana, basada únicamente en el ensayo y error, y que, por lo general, genera precisamente esa secuencia. Además, los políticos y quienes administran nuestros servicios sanitarios han recibido la idea con enorme alegría. Ellos ya sospechaban desde hacía mucho tiempo que los médicos eran totalmente acríticos y ahora lo veían reflejado por escrito. La medicina basada en la evidencia llegó como un regalo de los dioses ya que, al menos en su opinión, su eficiencia implícita debe reflejarse obligatoriamente en un ahorro de costes.

Sin embargo, el concepto de los ensayos clínicos controlados y la medicina basada en la evidencia no es nuevo. Las crónicas históricas reflejan que Federico II, emperador de los romanos y rey de Sicilia y de Jerusalén, que vivió entre los años 1192 y 1250 de nuestra era, y que estaba interesado en los efectos del ejercicio sobre la digestión, eligió a dos caballeros y les dio comidas idénticas. Uno fue enviado a cazar y al otro se le ordenó que durmiera. Después de varias horas, mató a ambos y examinó el contenido de sus tubos digestivos; la digestión había avanzado más en el estómago del caballero que había dormido. En el siglo XVII, Jan Baptista van Helmont, médico y filósofo, se volvió escéptico sobre la práctica de la sangría y propuso lo que casi seguro fue el primer ensayo clínico que incluyó a un gran número de participantes, con asignación aleatoria y análisis estadístico. Dicho estudio consistió en elegir a 200-500 personas pobres, dividirlas en dos grupos de forma aleatoria y evitar que en uno se realizase una sangría mientras que en el otro se permitía la realización de esta técnica en la medida en que sus colegas considerasen apropiado. El número de funerales en cada grupo se utilizaría para evaluar la eficacia de la sangría. La historia no registra por qué este magnífico experimento nunca se llevó a cabo.

Si hay que fijar el origen de la medicina científica moderna, se puede decir que fue en París en el siglo XIX, gracias al trabajo y las enseñanzas de Pierre Charles

Alexandre Louis. Louis introdujo el análisis estadístico para la evaluación del tratamiento médico y, de paso, demostró que la sangría era un método inútil de tratamiento, aunque esto no modificó los hábitos de los médicos de la época ni durante muchos años después. A pesar de este trabajo pionero, pocos clínicos a ambos lados del Atlántico instaron a que se adoptasen los ensayos sobre los resultados clínicos, aunque el genetista Ronald Fisher enunció los principios del diseño experimental basado en cifras en la década de 1920. Este campo sólo comenzó a tener un gran impacto en la práctica clínica después de la Segunda Guerra Mundial, gracias a los trabajos pioneros de Sir Austin Bradford Hill y de los epidemiólogos británicos que lo siguieron, sobre todo Richard Doll y Archie Cochrane.

Sin embargo, aunque la idea de la medicina basada en la evidencia no es nueva, los discípulos modernos, como David Sackett y sus colegas, están realizando una gran aportación a la práctica clínica, no sólo por la popularización de la idea sino por persuadir a los clínicos de que no es un tema académico árido, sino más bien una forma de pensar que debería estar presente en todos los aspectos de la práctica médica. Aunque gran parte de ella se basa en megaensayos y metaanálisis, también debería utilizarse para influir en casi todas las actuaciones médicas. Después de todo, la profesión médica ha sufrido un lavado de cerebro durante años por parte de los examinadores de las facultades de medicina y de los colegios profesionales para creer que sólo hay una manera de explorar a un paciente. Nuestros rituales a la cabecera del paciente podrían llevarse a cabo con una mayor evaluación crítica, al igual que nuestras operaciones y pautas farmacológicas; esto también es cierto para casi todos los aspectos de la medicina.

A medida que la práctica clínica exija cada vez más dedicación y el tiempo para la lectura y la reflexión se vuelva aún más valioso, la capacidad para examinar la literatura médica y, en el futuro, para familiarizarse con el conocimiento de la mejor práctica a partir de los sistemas de comunicación modernos será una habilidad esencial para los médicos. En este intenso libro, Trisha Greenhalgh ofrece una excelente aproximación a cómo hacer el mejor uso de la literatura médica y los beneficios de la medicina basada en la evidencia. Debería resultar igual de atractivo para los estudiantes de medicina de primer año que para los especialistas avezados y merece ser leído por un público muy amplio.

Con el paso de los años, el privilegio de ser invitado a escribir un prólogo para un libro de algún ex alumno se convierte en algo relativamente frecuente. Trisha Greenhalgh era el tipo de estudiante de medicina que nunca dejaba que sus profesores se saliesen por la tangente y esta actitud inquisitiva parece haber florecido en los últimos años; el lector tiene en sus manos un libro espléndido y oportuno, al cual deseo todo el éxito que se merece. Después de todo, el concepto de la medicina basada en la evidencia no es más que el estado de ánimo que cualquier profesor clínico espera desarrollar en sus alumnos; el enfoque escéptico, pero constructivo, de la Dra. Greenhalgh respecto a la literatura médica sugiere que es posible lograr un resultado tan satisfactorio al menos una vez en la vida de un profesor de medicina.

DJ Weatherall
Oxford

Prefacio a la primera edición: ¿es necesario leer este libro?

Este libro está dirigido a cualquier persona, tanto si tiene un título en medicina como si no, que desee orientarse en la literatura médica, evaluar la validez científica y la relevancia práctica de los artículos que se encuentren y, en su caso, aplicar los resultados en la práctica. Estas habilidades constituyen los fundamentos de la medicina basada en la evidencia.

Espero que este libro sea de ayuda para leer e interpretar mejor los artículos médicos. Es mi deseo, también, transmitir un mensaje adicional, que expondré a continuación. Muchas de las descripciones ofrecidas por los cínicos de lo que es la medicina basada en la evidencia (la glorificación de las cosas que pueden medirse sin tener en cuenta la utilidad o exactitud de lo que se mide, la aceptación acrítica de los datos numéricos publicados, la preparación de guías que abarcan todo por autoproclamados «expertos» que están alejados de la medicina real, la degradación de la libertad clínica por la imposición de protocolos clínicos rígidos y dogmáticos, y el exceso de confianza en análisis económicos simplistas, inapropiados y, a menudo, incorrectos) en realidad son críticas *contra* aquello que el movimiento de la medicina basada en la evidencia combate y no lo que representa.

Sin embargo, no quiero ser considerada una evangelista de la medicina basada en la evidencia. Creo que la ciencia de la búsqueda, evaluación y aplicación de los resultados de la investigación médica puede (y a menudo lo hace) lograr que la asistencia del paciente sea más objetiva, más lógica y más coste-efectiva. Si yo no creyese en esto, no dedicaría tanto tiempo a enseñarla e intentar practicarla como médico general. Sin embargo, creo que cuando se aplica en vacío (es decir, sin sentido común y sin tener en cuenta las circunstancias y prioridades de la persona a la cual se ofrece tratamiento o la compleja naturaleza de la práctica clínica y la elaboración de políticas), la toma de decisiones «basada en la evidencia» es un proceso reduccionista que puede ser verdaderamente perjudicial.

Por último, el lector debe tener en cuenta que no soy epidemióloga ni estadística, sino una persona que lee artículos y que ha desarrollado un sistema pragmático (y, en ocasiones, poco convencional) para evaluar su calidad. Animo a quien desee profundizar en los temas epidemiológicos y estadísticos incluidos en este libro a consultar textos específicos, cuyas referencias se encuentran al final de cada capítulo.

Trisha Greenhalgh

Prefacio a la quinta edición

Cuando escribí este libro en 1996, la medicina basada en la evidencia era una especie de punta de iceberg. Un puñado de académicos (entre los que me cuento) ya estaban entusiasmados y habían comenzado a impartir cursos de «formación para formadores» con el fin de difundir lo que considerábamos un enfoque muy lógico y sistemático de la práctica clínica. Otros (sin duda, la mayoría de los médicos) estaban convencidos de que se trataba de una moda pasajera que tenía una importancia limitada y que nunca arraigaría. Escribí *Cómo leer un artículo científico* por dos razones. En primer lugar, los estudiantes de mis propios cursos me solicitaban una introducción sencilla a los principios presentados en lo que por aquel entonces se denominaba el *gran libro rojo* de Dave Sackett (Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. Londres: Little, Brown & Co., 1991), un volumen excepcional e inspirador que ya iba por su cuarta reimpresión, pero que algunos principiantes consideraban de difícil lectura. En segundo lugar, yo tenía claro que muchos de los críticos de la medicina basada en la evidencia no entendían realmente lo que estaban rechazando y que, hasta que lo hicieran, no podría comenzar un debate serio sobre el lugar político, ideológico y pedagógico de la medicina basada en la evidencia como disciplina.

No puedo negar que estoy encantada de que *Cómo leer un artículo científico* se haya convertido en una lectura estándar en muchas facultades de medicina y enfermería, y que hasta ahora haya sido traducido al francés, alemán, italiano, español, portugués, chino, polaco, japonés, checo y ruso. También me alegro de que lo que hasta hace muy poco era un tema marginal del ámbito académico haya sido integrado adecuada y verdaderamente en la práctica clínica. En Reino Unido, por ejemplo, actualmente es un requisito contractual que todos los médicos, enfermeras y farmacéuticos ejerzan (y que los administradores gestionen) de acuerdo con la mejor evidencia de la investigación.

En los 18 años transcurridos desde que se publicó la primera edición de este libro, la medicina basada en la evidencia ha sufrido altibajos de popularidad. En la actualidad, cientos de libros y decenas de miles de artículos de revistas ofrecen diferentes ángulos sobre los «fundamentos de la MBE» que se describen brevemente en los capítulos de este libro. Un número cada vez mayor de estas

fuentes subrayan las verdaderas limitaciones de la medicina basada en la evidencia en ciertos contextos. Otros contemplan la medicina basada en la evidencia como un movimiento social —un «carro» que se puso en marcha en un momento (la década de 1990) y un lugar (Norteamérica) determinados y que se extendió muy deprisa con todo tipo de efectos en cascada para determinados grupos de interés—.

Durante la preparación de esta quinta edición he vuelto a tomar la decisión de no cambiar demasiado, aparte de actualizar los ejemplos y las listas de referencias, ya que todavía queda sitio en las librerías para un texto introductorio sencillo. En la edición previa añadí dos capítulos nuevos (sobre la mejora de la calidad y las intervenciones complejas) y en esta última he añadido otros dos, uno sobre la aplicación de la medicina basada en la evidencia con los pacientes (la ciencia de la toma de decisiones compartida) y otro sobre las críticas frecuentes contra la MBE y las respuestas. Como siempre, doy la bienvenida a cualquier comentario que ayude a hacer el texto más preciso, legible y práctico.

Trisha Greenhalgh
Enero de 2014

Agradecimientos

No soy en absoluto una experta en todos los temas tratados en este libro (en particular, se me dan muy mal los números) y quiero expresar mi agradecimiento a las personas que enumero a continuación por la ayuda que me han prestado durante la elaboración de este libro. En cualquier caso, soy la autora principal de cada capítulo y la responsabilidad de cualquier inexactitud es sólo mía.

1. A los profesores Sir Andy Haines y David Sackett, que me introdujeron en el tema de la medicina basada en la evidencia y me animaron a escribir sobre ello.
2. A la difunta Dra. Anna Donald, que amplió mi visión mediante valiosas conversaciones acerca de las implicaciones y las incertidumbres de esta disciplina evolutiva.
3. A Jeanette Buckingham, de la Universidad de Alberta, Canadá, por sus preciadas aportaciones al capítulo 2.
4. A los numerosos asesores expertos y revisores que han participado directamente en esta nueva edición o que me han aconsejado en ediciones anteriores.
5. A los incontables lectores, demasiados para mencionarlos individualmente, que han dedicado parte de su tiempo a escribir y señalar tanto los errores tipográficos como los de contenido en ediciones anteriores. Gracias a sus contribuciones, he aprendido mucho (sobre todo acerca de la estadística) y el libro ha mejorado en muchos aspectos. Algunos de los primeros críticos de *Cómo leer un artículo científico* han colaborado posteriormente conmigo en mis cursos docentes de la práctica basada en la evidencia; varios de ellos han sido coautores de artículos o de capítulos de libros conmigo, y uno o dos se han convertido en amigos personales.
6. A los autores y editores de artículos que me autorizaron a reproducir figuras o tablas. Los detalles se indican en el texto.
7. A mis seguidores en Twitter, que propusieron numerosas ideas, críticas constructivas y respuestas a mis sugerencias cuando estaba preparando la quinta edición de este libro. Por cierto, Twitter debería considerarse una fuente de información basada en la evidencia. Los lectores que lo deseen pueden seguirme en @trishgreenhalgh y también pueden seguir a la Cochrane Collaboration en @cochrancollab, a Ben Goldacre en @bengoldacre, a Carl Heneghan del Oxford

Centre for Evidence Based Medicine en @cebmblog y al National Institute for Health and Care Excellence en @nicecomms.

Gracias también a mi esposo, el Dr. Fraser Macfarlane, por su constante apoyo a mi trabajo académico y de escritora. Mis hijos Rob y Al aún eran muy pequeños cuando estaba escribiendo la primera edición de este libro. Me llena de orgullo que ahora hayan leído el libro, hayan aplicado sus mensajes en sus incipientes carreras científicas (uno de ellos en medicina) y me hayan aportado sugerencias sobre cómo mejorarlo.

Capítulo 1 ¿Para qué molestarse en leer artículos científicos?

¿La «medicina basada en la evidencia» consiste simplemente en «leer artículos en revistas médicas»?

La medicina basada en evidencia (MBE) es mucho más que limitarse a leer artículos científicos. Según la definición más citada, es «el uso consciente, explícito y sensato de la mejor evidencia actual en la toma de decisiones sobre la asistencia de los pacientes concretos»[1]. En mi opinión, esta definición es muy útil, pero no incluye un aspecto que considero muy importante sobre el tema, y que es el uso de las matemáticas. Aunque no sepa casi nada sobre MBE, es probable que esté al tanto de que los números y las razones *(ratios)* están omnipresentes en ella. Anna Donald y yo decidimos ser consecuentes con esto en nuestra vertiente docente y propusimos esta definición alternativa:

> *La medicina basada en la evidencia es el uso de estimaciones matemáticas del riesgo de beneficios y perjuicios, derivadas de la investigación de alta calidad en muestras poblacionales, para documentar la toma de decisiones clínicas en el diagnóstico, la investigación o el tratamiento de los pacientes concretos.*

Por lo tanto, la característica definitoria de la MBE es el uso de cifras derivadas de la investigación sobre las *poblaciones* para informar las decisiones sobre las *personas*. Esto, por supuesto, suscita la pregunta: «¿qué es la investigación?», para la cual una respuesta razonablemente precisa podría ser: «una indagación centrada y sistemática dirigida a generar nuevos conocimientos». En capítulos posteriores explicaré cómo esta definición puede ayudar a distinguir la investigación genuina (que debería utilizarse como fuente de información para la práctica médica) de los intentos de mala calidad de aficionados bienintencionados (que se deben ignorar educadamente).

Por consiguiente, si se sigue un enfoque basado en la evidencia para la toma de decisiones clínicas, diversas cuestiones de todo tipo relativas a los pacientes (o, si se trabaja en el ámbito de la salud pública, cuestiones relativas a grupos de

personas) llevarán a plantearse preguntas sobre la evidencia científica, a buscar respuestas a estas preguntas de manera sistemática y a modificar la práctica en consecuencia.

Es posible plantearse preguntas, por ejemplo, sobre los síntomas de un paciente («¿cuál es la probabilidad de que un varón de 34 años con dolor torácico unilateral izquierdo tenga un problema cardíaco grave? Y, si lo tiene, ¿se observará en un ECG de reposo?»), sobre los signos físicos o diagnósticos («en un parto sin otras complicaciones, ¿la presencia de meconio [indicativa de una defecación fetal] en el líquido amniótico sugiere un deterioro significativo del estado fisiológico del feto?»), sobre el pronóstico de una enfermedad («si una niña de 2 años previamente sana tiene una crisis comicial breve asociada con fiebre, ¿cuál es la probabilidad de que desarrolle posteriormente epilepsia?»), sobre el tratamiento («en los pacientes con un síndrome coronario agudo [ataque cardíaco], los riesgos asociados con los fármacos trombolíticos [eliminadores del coágulo] se ven superados por los beneficios, con independencia de su edad, sexo y origen étnico?»), sobre la rentabilidad («¿está justificado el coste de este nuevo fármaco contra el cáncer en comparación con otras formas de gastar unos recursos sanitarios limitados?»), sobre las preferencias de los pacientes («en una mujer de 87 años con fibrilación auricular intermitente y un ataque isquémico transitorio reciente, ¿los inconvenientes del tratamiento con warfarina son mayores que los riesgos de no tomarla?») y sobre otros aspectos diversos referentes a la salud y los servicios sanitarios.

El profesor Sackett, en el editorial de apertura del primer número de la revista *Evidence-Based Medicine,* resumió los pasos esenciales en la ciencia emergente de la MBE[2]:

1. Convertir nuestras necesidades de información en preguntas que puedan obtener respuestas (es decir, formular el problema).
2. Localizar, del modo más eficiente, la mejor evidencia con la que responder a estas preguntas, que puede provenir de la exploración física, las pruebas de laboratorio, la literatura publicada o de otras fuentes.
3. Realizar una evaluación crítica de la evidencia (es decir, ponderarla) para determinar su validez (cercanía a la verdad) y utilidad (aplicabilidad clínica).
4. Implementar los resultados de esta evaluación en nuestra práctica clínica.
5. Evaluar nuestro rendimiento.

Por lo tanto, la MBE no sólo requiere leer artículos científicos, sino leer los artículos *correctos* en el momento adecuado y después modificar la conducta (y, lo que a menudo es más difícil, influir en la conducta de otras personas) a la luz de lo que se ha encontrado. Me preocupa que los cursos prácticos sobre MBE se centran demasiado a menudo en la tercera de estas cinco etapas (evaluación crítica) dejando de lado todas las demás. Sin embargo, si se ha planteado la pregunta errónea o si se buscan las respuestas en las fuentes equivocadas, también es posible que no se haya leído ningún artículo científico. Asimismo, toda la formación sobre técnicas de búsqueda y evaluación crítica será en vano si no se

pone al menos el mismo esfuerzo en la aplicación de la evidencia válida y en la medición del progreso hacia los objetivos como el que se dedica a la lectura de los artículos. Hace unos años añadí tres etapas más al modelo de cinco etapas de Sackett para incorporar la perspectiva del paciente: las ocho etapas resultantes, que he denominado *lista de comprobación sensible al contexto para la práctica basada en la evidencia,* se muestran en el apéndice 1[3].

Si me tomase el título de este libro al pie de la letra, estos aspectos más amplios de la MBE ni siquiera deberían haberse mencionado aquí, pero correría el riesgo de que los lectores exigieran que se les devolviese su dinero si hubiese omitido la sección final de este capítulo (Antes de empezar: formule el problema), el capítulo 2 (Búsquedas en la literatura), el capítulo 15 (Aplicación de la práctica basada en la evidencia) y el capítulo 16 (Aplicación de la evidencia con los pacientes). Los capítulos 3-14 describen el paso tres del proceso de MBE: valoración crítica, es decir, lo que debemos hacer cuando tenemos el artículo científico delante de nosotros. El capítulo 16 presenta los aspectos que suelen criticarse a la MBE.

Por cierto, los lectores con conocimientos de informática que quieran indagar sobre la MBE en internet pueden navegar por las páginas web que aparecen en el cuadro 1.1. Si la informática no es lo suyo, no se preocupe por el momento, pero añada a su lista de tareas pendientes una anotación para recordarle que debe aprender/utilizar los recursos de internet. No se preocupe cuando descubra que hay más de 1.000 páginas web dedicadas a la MBE: todas ellas ofrecen material muy similar y no es necesario visitarlas todas.

Cuadro 1.1 Recursos de internet para la medicina basada en la evidencia

Oxford Centre for Evidence-Based Medicine: un sitio web excelente de Oxford, Reino Unido, que contiene gran cantidad de recursos y enlaces sobre la MBE. http://cebm.net

National Institute for Health and Care Excellence: este sitio web de Reino Unido, que también es popular fuera de este país, tiene enlaces a guías clínicas basadas en la evidencia y revisiones temáticas. http://www.nice.org.uk/

National Health Service (NHS) Centre for Reviews and Dissemination: este sitio, que permite descargar revisiones de alta calidad basadas en la evidencia, forma parte del National Institute for Health Research de Reino Unido y constituye un buen punto de partida para buscar evidencia sobre cuestiones complejas como: «¿qué se debe hacer con la obesidad?». http://www.york.ac.uk/inst/crd/

Clinical Evidence: un manual en línea sobre la mejor evidencia para decisiones clínicas como: «¿cuál es el mejor tratamiento actual para la fibrilación auricular?». Este sitio está administrado por BMJ Publishing Group. http://clinicalevidence.bmj.com

¿Por qué hay quien se queja cuando se menciona la medicina basada en la evidencia?

Es posible que quienes critican la MBE la definan como «la tendencia de un grupo de académicos médicos jóvenes, seguros y con sólidos conocimientos matemáticos a menospreciar la práctica de los clínicos experimentados utilizando una combinación de jerga epidemiológica y prestidigitación estadística» o «el argumento, presentado habitualmente con un celo casi evangelizador, de que ningún médico, enfermera, solicitante de servicios sanitarios o político debería emprender ninguna acción a menos que se hayan publicado los resultados de varios ensayos de investigación a gran escala y costosos, y que hayan sido aprobados por un comité de expertos».

El resentimiento que existe entre algunos profesionales sanitarios contra el movimiento de la MBE es principalmente una reacción a la implicación de que los médicos (y enfermeras, matronas, fisioterapeutas y otros profesionales sanitarios) eran analfabetos funcionales hasta que se les mostró la luz y que los pocos que no eran analfabetos ignoraban deliberadamente la evidencia médica publicada. Cualquiera que trabaje cara a cara con pacientes conoce con cuánta frecuencia surge la necesidad de buscar nueva información antes de tomar una decisión clínica. Los médicos han pasado gran parte de su tiempo en las bibliotecas desde que éstas se inventaron. En general, no se prescribe un nuevo medicamento a los pacientes sin contar con evidencia de su probable eficacia. Al margen de otras cuestiones, ese uso fuera de las indicaciones autorizadas de los fármacos es, en sentido estricto, ilegal. Seguramente todos hemos practicado MBE durante años, salvo cuando hacíamos un uso deliberado del «efecto placebo» por unas razones médicas justificadas, o cuando estábamos enfermos, hiperestresados o siendo conscientemente perezosos.

En realidad, no todos los profesionales lo han hecho. Se han llevado a cabo varios estudios sobre la conducta de los médicos, enfermeras y profesionales relacionados. En la década de 1970, en EE.UU., se estimaba que sólo alrededor del 10-20% de todas las tecnologías de la salud disponibles en ese momento (es decir, fármacos, procedimientos, operaciones, etc.) estaban basadas en la evidencia; esa cifra mejoró al 21% en 1990, según las estadísticas estadounidenses oficiales[4]. Los estudios de las intervenciones que se ofrecen a series consecutivas de pacientes sugieren que el 60-90% de las decisiones clínicas, dependiendo de la especialidad, estaban «basadas en la evidencia»[5]. Sin embargo, como he argumentado en otra parte del libro, estos estudios tenían limitaciones metodológicas[3]. Entre otras cosas, se llevaron a cabo en unidades especializadas y analizaron la práctica de expertos mundiales en MBE; por lo tanto, resulta prácticamente imposible que las cifras obtenidas se puedan generalizar más allá de su entorno inmediato (v. sección «¿Qué pacientes incluye el estudio?»). Lo más probable es que todavía estemos infravalorando a nuestros pacientes la mayor parte del tiempo.

En un estudio reciente a gran escala, realizado por un equipo australiano, se evaluó a 1.000 pacientes tratados por las 22 afecciones vistas con más frecuencia

en un centro de atención primaria. Los investigadores observaron que, mientras que el 90% de los pacientes recibieron atención basada en la evidencia para la cardiopatía isquémica, esta cifra sólo era del 13% para la dependencia del alcohol[6]. Por otra parte, el grado en el que cualquier médico proporcionaba una asistencia basada en la evidencia variaba en la muestra del 32% al 86% de las veces. Estos resultados sugieren que aún hay mucho margen para mejorar.

A continuación, echaremos un vistazo a los distintos enfoques que los profesionales sanitarios emplean para tomar sus decisiones en la vida real y que constituyen ejemplos de lo que *no es* MBE.

Toma de decisiones basada en casos anecdóticos

En mi época de estudiante de medicina, de vez en cuando me unía a la comitiva de un profesor distinguido en sus rondas de planta diarias. Cuando veíamos a un nuevo paciente, él preguntaba sobre sus síntomas, se volvía al grupo de residentes que se congregaban alrededor de la cama y contaba la historia de un paciente semejante al que había atendido unos años antes. «Ah, sí. Recuerdo que le dimos tal y tal cosa, y después de eso mejoró.» Mostraba una postura cínica, a menudo con razón, sobre nuevos fármacos y tecnologías, y tenía una perspicacia clínica insuperable. Sin embargo, había necesitado 40 años para acumular su experiencia y el tratado de medicina más voluminoso que existe –el que contiene el conjunto de casos a los que nunca atendió– quedó fuera de su alcance para siempre.

Los casos anecdóticos (relatos de casos concretos) ocupan un lugar importante en la práctica clínica[7]. Los psicólogos han demostrado que los estudiantes adquieren las habilidades de la medicina, la enfermería y otras profesiones sanitarias memorizando lo que les pasaba a pacientes concretos y cuál fue su evolución, en forma de historias o «guiones de enfermedad». Las historias sobre los pacientes son la unidad de análisis (es decir, lo que estudiamos) en las sesiones clínicas y las clases. Los médicos obtienen información crucial a partir del relato que hacen los pacientes sobre su enfermedad y, lo que quizá sea más importante, así averiguan lo que estar enfermo *significa* para el paciente. Los médicos y enfermeras experimentados se basan en los «guiones de enfermedad» acumulados de todos sus pacientes previos a la hora de tratar a los pacientes posteriores. Pero eso no significa simplemente hacer con el paciente B lo mismo que se hizo con el paciente A si el tratamiento funcionó y hacer exactamente lo contrario si fracasó.

Los peligros de la toma de decisiones basada en casos anecdóticos quedan bien ilustrados cuando se considera la relación riesgo-beneficio de los fármacos. Durante mi primer embarazo tuve vómitos intensos y me prescribieron proclorperazina, un fármaco antiemético. A los pocos minutos, me apareció un espasmo neurológico incontrolable y muy desagradable. Dos días más tarde, me había recuperado por completo de esta reacción idiosincrásica, pero nunca he recetado el fármaco desde entonces, a pesar de que la prevalencia estimada de reacciones neurológicas a la proclorperazina es de tan sólo de uno entre varios miles de casos. Por el contrario, es tentador desechar la posibilidad de que se produzcan efectos adversos poco frecuentes, pero potencialmente graves, de fármacos muy

conocidos (como la trombosis debida a la píldora anticonceptiva) cuando nunca se han observado estos problemas en uno mismo o en los propios pacientes.

Los médicos no seríamos humanos si ignorásemos nuestras experiencias clínicas personales, pero sería mejor basar nuestras decisiones en la experiencia colectiva de miles de médicos que tratan a millones de pacientes, en lugar de en lo que hemos visto y sentido cada uno de nosotros. En el capítulo 5 (Estadística para no estadísticos) se describen algunos de los métodos más objetivos, como el número que es necesario tratar (NNT), para decidir si es probable que un fármaco particular (u otra intervención) cause un beneficio o perjuicio significativo a un paciente.

Cuando el movimiento MBE estaba todavía en pañales, Sackett hizo hincapié en que la práctica basada en la evidencia no era una amenaza para la experiencia o juicio clínico de la antigua escuela[1]. La cuestión de *cómo* pueden actuar los médicos para que su práctica sea a la vez «basada en la evidencia» (es decir, documentando sistemáticamente sus decisiones con la evidencia basada en la investigación) y «basada en la narrativa» (es decir, integrando todos sus casos clínicos concretos acumulados y tratando el problema de cada paciente como una única historia clínica en lugar de cómo «un caso de enfermedad X») es difícil de abordar desde un punto de vista filosófico y queda fuera del alcance de este libro. Los lectores que tengan interés pueden consultar dos artículos que he escrito sobre este tema[8,9].

Toma de decisiones basada en recortes de prensa

Durante los primeros 10 años después de graduarme, tenía un archivo creciente de artículos que arrancaba de mis semanarios médicos antes de tirar a la papelera los contenidos menos interesantes. Si un artículo o un editorial parecía tener algo nuevo que decir, yo modificaba conscientemente mi práctica clínica de acuerdo con sus conclusiones. Si un artículo afirmaba que todos los niños con sospecha de infecciones del tracto urinario debían ser remitidos para realizar exploraciones renales que descartasen anomalías congénitas, yo enviaba a todos los pacientes menores de 16 años con síntomas urinarios a someterse a estudios especializados. Como la recomendación se había publicado y era reciente, no había dudas de que debía reemplazar la práctica estándar previa, que en este caso consistía en remitir únicamente al pequeño porcentaje de esos niños que mostrasen características «atípicas»[10].

Esta estrategia de toma de decisiones clínicas es muy común todavía. ¿Cuántos médicos conoce que justifiquen su estrategia frente a un problema clínico particular citando la sección de resultados de un único estudio publicado aunque no sepan nada en absoluto acerca de los métodos utilizados para obtener esos resultados? ¿Se trataba de un ensayo aleatorizado y controlado? (v. sección «Estudios transversales»). ¿Cuántos pacientes, de qué edad, sexo y gravedad de la enfermedad incluía el estudio? (v. sección «¿Qué pacientes incluye el estudio?»). ¿Cuántos abandonaron el estudio y por qué? (v. sección «¿Se abordaron las cuestiones estadísticas preliminares?»). ¿Con qué criterio se consideró que los pacientes estaban curados? (v. sección «Criterios de valoración indirectos»). Si los

resultados del estudio parecían contradecir los de otros investigadores, ¿qué se hizo para intentar validarlos (confirmarlos) y replicarlos (repetirlos)? (v. sección «Diez preguntas que deben plantearse sobre un artículo que pretende validar una prueba diagnóstica o de cribado»). ¿Las pruebas estadísticas que supuestamente demostraban la opinión de los autores fueron elegidas de forma apropiada y realizadas correctamente? (v. cap. 5). ¿Se tuvo en cuenta sistemáticamente el punto de vista del paciente y se incorporó mediante una herramienta para la toma de decisiones compartida? (v. cap. 16). Los médicos (y enfermeras, matronas, gestores médicos, psicólogos, estudiantes de medicina y defensores de los derechos de los consumidores) a quienes les gusta citar los resultados de los estudios de investigación médica tienen la responsabilidad de comprobar que primero repasan una lista de comprobación de preguntas como éstas (se incluyen otras adicionales en el apéndice 1).

Toma de decisiones basadas en un consenso de expertos

Cuando escribí la primera edición de este libro a mediados de la década de 1990, el tipo más frecuente de guías clínicas era lo que se denomina «declaración de consenso», consistente en el fruto del duro trabajo de alrededor de una docena de eminentes expertos que habían pasado un fin de semana encerrados en un hotel de lujo, por lo general pagados por una compañía farmacéutica. Las guías clínicas derivadas de este consenso de expertos a menudo se elaboraban a partir de publicaciones médicas gratuitas (revistas médicas y otras «hojas informativas» gratuitas patrocinadas directa o indirectamente por la industria farmacéutica), como folletos de bolsillo repletos de recomendaciones resumidas y guías terapéuticas rápidas. Sin embargo, ¿quién dice que los consejos dados en un conjunto de guías clínicas, un editorial impactante o una revisión con una extensa bibliografía son correctos?

La profesora Mulrow[11], pionera de la revisión sistemática (v. cap. 9), demostró hace unos años que los expertos en un campo clínico particular tienen *menos* probabilidades de realizar una revisión objetiva de toda la evidencia disponible que alguien que no sea experto que aborda la literatura con ojos imparciales. En casos extremos, una «opinión experta» puede consistir simplemente en los malos hábitos y los recortes de prensa personales atesorados durante la vida de un médico de edad avanzada, y un grupo de estos expertos únicamente multiplicará los puntos de vista equivocados de cualquiera de ellos. En la tabla 1.1 se presentan ejemplos de prácticas que en su momento fueron ampliamente aceptadas como buena práctica clínica (y que habrían figurado en las guías clínicas del consenso de expertos de su época), pero que posteriormente han sido desacreditadas por ensayos clínicos de alta calidad.

En el capítulo 9 se presenta una lista de comprobación para evaluar si una «revisión sistemática de la evidencia» elaborada con el fin de respaldar las recomendaciones para la práctica o la formulación de políticas realmente merece la pena, y en el capítulo 10 se describe lo perjudicial que puede ser la aplicación de guías clínicas que no están basadas en la evidencia. El hecho de que en la

Tabla 1.1 Ejemplos de prácticas nocivas que contaron en su momento con el respaldo de la «opinión de expertos»

Época aproximada	Práctica clínica aceptada por los expertos de la época	Año en que se demostró que era perjudicial	Impacto sobre la práctica clínica
Desde el año 500 a.C.	Sangrías (para casi cualquier enfermedad aguda)	1820[a]	Las sangrías se interrumpieron alrededor de 1910
1957	Talidomida para las náuseas matutinas en el primer trimestre del embarazo, lo que provocó el nacimiento de más de 8.000 bebés con graves malformaciones en todo el mundo	1960	Los efectos teratógenos de este fármaco fueron tan dramáticos que la talidomida se retiró enseguida cuando apareció la primera publicación de un caso
Al menos desde 1900	Reposo en cama para la lumbalgia aguda	1986	Muchos médicos aún aconsejan «reposo» a los pacientes con lumbalgia
Década de 1960	Benzodiazepinas (p. ej., diazepam) para la ansiedad leve y el insomnio, comercializadas inicialmente como no adictivas, pero posteriormente se demostró que causaban una dependencia intensa y síntomas de abstinencia	1975	La prescripción de benzodiazepinas para estas indicaciones disminuyó en la década de 1990
Década de 1970	Lidocaína intravenosa en el infarto de miocardio agudo, como profilaxis antiarrítmica. Posteriormente se demostró que no proporcionaba beneficios globales y que, en algunos casos, provocaba arritmias mortales	1974	La lidocaína siguió administrándose de forma rutinaria hasta mediados de la década de 1980
Finales de la década de 1990	Inhibidores de la COX-2 (una nueva clase de antiinflamatorios no esteroideos), introducidos para el tratamiento de la artritis. Posteriormente se demostró que aumentaban el riesgo de ataque cardíaco y de ictus	2004	Los inhibidores de la COX-2 para el dolor se retiraron rápidamente después de varios casos legales con gran repercusión en EE.UU., aunque actualmente se están evaluando nuevos usos en el tratamiento antioncológico (donde los beneficios pueden superar a los riesgos)

[a]Curiosamente, las sangrías fueron probablemente la primera práctica para la que se sugirió un ensayo clínico controlado y aleatorizado. El médico Van Helmont planteó este reto a sus colegas en 1662: «*Tomemos a 200-500 personas con fiebre. Hagamos un sorteo para que la mitad sean tratados por mí y la otra mitad por vosotros. Yo los curaré sin sangrías, pero vosotros lo haréis a vuestra manera, para ver cuántos pacientes fallecen en cada grupo*»[12]. Quiero dar las gracias a Matthias Egger por haberme prestado este ejemplo.

actualidad casi no existan guías derivadas de consensos de expertos es uno de los logros principales de la MBE.

Toma de decisiones basada en la minimización de costes

La prensa popular tiende a horrorizarse cuando descubre que no se ha administrado un tratamiento a un paciente debido a su coste. Los gerentes, políticos y, cada vez más, los médicos deben saber que serán puestos en la picota cuando un niño con un cáncer poco frecuente no sea remitido a una unidad especializada en EE.UU. o cuando se niegue a una frágil anciana un fármaco para detener su pérdida de visión por degeneración macular. Sin embargo, en el mundo real, cualquier asistencia sanitaria se presta a partir de un presupuesto limitado y cada vez hay una mayor conciencia de que las decisiones clínicas deben tener en cuenta los costes económicos de una determinada intervención. Como se expone en el capítulo 11, la toma de decisiones clínicas basada *exclusivamente* en los costes («minimización de costes»: elección de la alternativa más barata sin tener en cuenta su eficacia) no suele estar justificada desde el punto de vista ético y tenemos derecho a expresar nuestra oposición cuando esto sucede.

Sin embargo, las intervenciones caras no deberían justificarse simplemente porque sean nuevas, porque deberían funcionar en teoría o porque la única alternativa sea no hacer nada, sino porque sea muy probable que salven la vida o que mejoren significativamente su calidad. Pero, ¿cómo es posible realizar una comparación comprensible entre los beneficios de una artroplastia total de cadera en un paciente de 75 años y los fármacos hipocolesterolemiantes en un varón de mediana edad o la evaluación de la infertilidad en una pareja de veinteañeros? En contra del sentido común, no hay un conjunto evidente de principios éticos o de herramientas analíticas que se puedan utilizar para ajustar unos recursos limitados a una demanda ilimitada. Como se verá en el capítulo 11, el tan denigrado concepto de años de vida ajustados por calidad o AVAC (QALY por su acrónimo en inglés) y otras unidades similares basadas en la utilidad son simplemente un intento de proporcionar cierta objetividad a la comparación ilógica, pero inevitable, entre peras y manzanas en el ámbito del sufrimiento humano. En Reino Unido, el National Institute for Health and Care Excellence (v. www.nice.org.uk) trata de desarrollar tanto guías clínicas basadas en la evidencia como una asignación equitativa de los recursos del NHS.

Existe una razón adicional por la cual algunas personas consideran que el término *medicina basada en la evidencia* es desagradable. En este capítulo se ha argumentado que la MBE consiste en afrontar el cambio, no en intentar saber todas las respuestas antes de empezar. Dicho de otro modo, no se trata tanto de lo que se haya leído en el pasado sino de lo que se hace para identificar y satisfacer las necesidades de aprendizaje continuo y para aplicar los conocimientos de un modo apropiado y homogéneo en nuevas situaciones clínicas. Los médicos que se habían formado con el estilo de la vieja escuela de no admitir nunca la ignorancia pueden tener dificultades para aceptar que existe un elemento importante de incertidumbre científica en prácticamente todas las situaciones clínicas, aunque en

la mayoría de los casos, el médico no logra identificar la incertidumbre o articularla en términos de una pregunta que se pueda contestar (v. sección siguiente). Los lectores interesados en la evidencia basada en la investigación sobre la (falta de) conducta crítica de los médicos pueden consultar una excelente revisión de Swinglehurst sobre el tema[13].

El hecho de que ninguno de nosotros, ni siquiera los más listos o los más experimentados, pueda responder a todas las preguntas que surgen en una situación clínica típica significa que el «experto» es más falible de lo que tradicionalmente se consideraba. Un enfoque basado en la evidencia de las sesiones clínicas puede dar la vuelta a la jerarquía médica tradicional cuando una enfermera o un médico residente elabore nueva evidencia que contradiga lo que el médico más experimentado enseñó a todos la semana pasada. Para algunos médicos mayores, el aprendizaje de las habilidades de evaluación crítica es el menor de sus problemas a la hora de adaptarse a un estilo de enseñanza basada en la evidencia.

Después de haber defendido la MBE contra todos los argumentos habituales planteados por los médicos, debo confesar que siento simpatía por muchos de los argumentos más sofisticados aducidos por filósofos y científicos sociales. Estos argumentos, que se resumen en el capítulo 17 (nuevo en esta edición), abordan la naturaleza del conocimiento y la cuestión de qué proporción de la medicina se basa realmente en decisiones, pero quiero pedir al lector que no salte hasta ese capítulo (que es una «lectura ardua» desde un punto de vista filosófico) hasta haber comprendido plenamente los argumentos básicos expuestos en los primeros capítulos de este libro o correrá el riesgo de sumirse en la confusión.

Antes de empezar: formule el problema

Cuando les pido a mis estudiantes de medicina que me escriban un ensayo sobre la hipertensión arterial, a menudo redactan unos párrafos largos, académicos y esencialmente correctos sobre lo que es la hipertensión arterial, cuáles son sus causas y qué opciones terapéuticas existen. El día que entregan sus ensayos, la mayoría de ellos saben mucho más acerca de hipertensión arterial que yo. En ese momento, saben que la hipertensión arterial es la causa más común de accidente cerebrovascular (ACV) y que la detección y el tratamiento de hipertensión arterial en todos los pacientes reducirían la incidencia de ACV casi un 50%. La mayoría de ellos son conscientes de que el ACV, aunque es devastador cuando se produce, es bastante infrecuente, y que los fármacos antihipertensivos tienen efectos secundarios, como cansancio, mareos, impotencia e incontinencia urinaria.

Sin embargo, cuando les planteo a mis alumnos una cuestión práctica como: «una paciente tiene episodios de mareo desde que toma esta medicación antihipertensiva y quiere dejar todos los medicamentos; ¿qué le aconsejaríais hacer?», a menudo se quedan desconcertados. Ellos son capaces de comprender la postura de la paciente, pero no son capaces de extraer de sus páginas densamente escritas lo único que la paciente debe saber. Esto recuerda a lo que Smith (parafraseando a T. S. Eliot) preguntó hace unos años en un editorial del BMJ: «¿dónde está la

sabiduría que hemos perdido en el conocimiento, y dónde el conocimiento que hemos perdido en la información?»[14].

Los médicos experimentados podrían pensar que son capaces de responder a la pregunta de esta paciente a partir de su propia experiencia personal. Como ya expuse en la sección anterior, pocos de ellos lo harían correctamente. E incluso si acertasen en esta ocasión, seguirían necesitando un sistema global para convertir el maremágnum de información sobre un paciente (un conjunto difuso de síntomas, signos físicos, resultados de pruebas e información de lo que sucedió a este paciente o a un paciente similar la última vez), los valores y preferencias particulares (utilidades) del paciente y otras cosas que podrían ser relevantes (una corazonada, un artículo recordado a medias, la opinión de un colega más experto o un párrafo descubierto por casualidad al hojear un libro de texto) en un breve resumen de cuál es el problema y qué elementos específicos adicionales de información se necesitan para resolverlo.

Sackett y cols., en un libro revisado después por Straus[15], han aportado su ayuda al diseccionar las partes de una pregunta clínica adecuada:

- En primer lugar, hay que definir con precisión *a quién* va dirigida la pregunta (es decir, se debe preguntar: «¿cómo describiría a un grupo de pacientes similares a éste?»).
- A continuación, hay que definir *qué* actuación se está considerando en este paciente o población (p. ej., un tratamiento farmacológico) y, en caso necesario, un método de comparación (p. ej., el placebo o el tratamiento estándar actual).
- Por último, se debe definir el *resultado* deseado (o no deseado) (p. ej., reducción de la mortalidad, mejor calidad de vida y ahorro global para el servicio de salud).

El segundo paso puede que no se refiera a un tratamiento farmacológico, operación quirúrgica u otra intervención. La actuación podría ser, por ejemplo, la exposición a un supuesto carcinógeno (algo que podría causar cáncer) o la detección de un criterio de valoración indirecto en un análisis de sangre o en otra prueba. (Un criterio de valoración indirecto, como se explica en la sección «Criterios de valoración indirectos», es algo que predice, o que supuestamente predice, el desarrollo o progresión posterior de la enfermedad. En realidad, hay muy pocas pruebas que sirvan de forma fiable como bolas de cristal para adivinar el futuro médico de los pacientes. La afirmación «el médico vio los resultados de la prueba y me dijo que me quedaban seis meses de vida» suele deberse a la mala memoria o a un médico irresponsable.) En ambos casos, el «resultado» sería el desarrollo de cáncer (o de otra enfermedad) varios años después. Sin embargo, en la mayoría de los problemas clínicos con pacientes concretos, la «actuación» consiste en una intervención específica iniciada por un profesional sanitario.

Por tanto, en el caso de nuestra paciente hipertensa, podríamos preguntar: «en una mujer de raza blanca de 68 años, con hipertensión esencial (es decir, presión arterial alta común), sin enfermedades asociadas ni antecedentes médicos significativos, cuya presión arterial actualmente es X/Y, ¿los beneficios de continuar el tratamiento con bendroflumetiazida (sobre todo, la reducción del riesgo de

ACV) son superiores a los inconvenientes?». Hay que tener en cuenta que, al formular la pregunta específica, ya va implícito que la paciente nunca ha tenido un ataque al corazón, un ACV o signos de alerta precoces tales como parálisis transitoria o pérdida de visión. Si hubiese presentado algo de esto, su riesgo de ACV subsiguiente sería mucho mayor y se debería haber modificado la ecuación riesgo-beneficio para reflejarlo.

Para responder a la pregunta que hemos planteado, debemos determinar no sólo el riesgo de ACV en la hipertensión no tratada, sino también la reducción probable de ese riesgo que podemos esperar con el tratamiento farmacológico. Esto supone, en realidad, reformular una pregunta más general (¿Los beneficios del tratamiento en este caso son mayores que los riesgos?) que deberíamos haber planteado antes de prescribir bendroflumetiazida a la paciente en primer lugar, y que todos los médicos deben preguntarse a sí mismos cada vez que echen mano de su talonario de recetas.

Debe recordarse que la alternativa de la paciente a seguir tomando ese medicamento en particular no consiste necesariamente en no tomar ningún medicamento en absoluto; puede haber otros fármacos con una eficacia equivalente, pero con efectos secundarios menos discapacitantes (como se expone en el cap. 6, demasiados ensayos clínicos de nuevos fármacos comparan el producto con placebo en lugar de con la mejor alternativa disponible), o tratamientos no médicos, como el ejercicio, la restricción de sal, la homeopatía o el yoga. No todos estos métodos serían de ayuda para la paciente o le resultarían aceptables, pero sería muy adecuado buscar evidencia sobre *si* podrían ayudarla, especialmente si ella solicitase probar uno o más de estos remedios.

Es probable que se puedan encontrar respuestas a algunas de estas preguntas en la literatura médica, y en el capítulo 2 se describe cómo buscar artículos pertinentes una vez que se ha formulado el problema. Sin embargo, antes de empezar, sugiero al lector que eche un último vistazo a nuestra paciente con hipertensión arterial. Para determinar sus prioridades personales (¿cómo valora ella una reducción del 10% de su riesgo de ACV en un período de 5 años comparada con la incapacidad para ir de compras hoy en día sin que nadie la acompañe?), habrá que enfocar el tema desde la perspectiva de la paciente, en vez de consultar con un especialista en hipertensión arterial o con la base de datos Medline. En el capítulo 16 se establecen algunos enfoques estructurados para hacerlo.

Ejercicio 1

1. Vuelva al cuarto párrafo de este capítulo, donde se presentan ejemplos de preguntas clínicas. Decida si cada una de ellas está enfocada correctamente en términos de:
   (a) El paciente o problema.
   (b) La actuación (intervención, marcador pronóstico, exposición).
   (c) La actuación comparativa, si procede.
   (d) El resultado clínico.

2. Ahora, responda a estas preguntas:
   (a) Un niño de 5 años lleva en tratamiento con dosis altas de esteroides tópicos para un cuadro de eccema intenso desde los 20 meses. La madre cree que los esteroides están impidiendo el crecimiento del niño y desea cambiar a un tratamiento homeopático. ¿Qué información necesita el dermatólogo para decidir (i) si la madre está en lo cierto acerca de los esteroides tópicos y (ii) si el tratamiento homeopático será de ayuda para este niño?
   (b) Una mujer embarazada de 9 semanas llama por teléfono a su médico de cabecera por dolor abdominal y hemorragia. Una ecografía anterior mostró que el embarazo no era ectópico. El médico decide que podría tratarse de un aborto y le dice que debe ir al hospital para una ecografía y, posiblemente, para realizar una operación de legrado uterino. La mujer se muestra reticente. ¿Qué información necesitan tanto la paciente como el médico para determinar si el ingreso hospitalario es médicamente necesario?
   (c) Un varón de 48 años consulta con un médico privado por lumbalgia. El médico administra una inyección de corticoides. Por desgracia, el paciente desarrolla una meningitis fúngica y fallece. ¿Qué información se necesita para determinar tanto los beneficios como los perjuicios potenciales de las inyecciones de corticoides en pacientes con lumbalgia, con el fin de aconsejarles sobre la relación riesgo-beneficio?

Bibliografía

1 Sackett DL, Rosenberg WM, Gray J, et al. Evidence based medicine: what it is and what it isn't. BMJ: British Medical Journal 1996;**312**(7023):71.
2 Sackett DL, Haynes RB. On the need for evidence-based medicine. Evidence Based Medicine 1995;**1**(1):4-5.
3 Greenhalgh T. Is my practice evidence-based? BMJ: British Medical Journal 1996;**313**(7063):957.
4 Dubinsky M, Ferguson JH. Analysis of theNational Institutes of Health Medicare coverage assessment. International Journal of Technology Assessment in Health Care 1990;**6**(03):480-8.
5 Sackett D, Ellis J, Mulligan I, et al. Inpatient general medicine is evidence based. The Lancet 1995;**346**(8972):407-10.
6 Runciman WB, Hunt TD, Hannaford NA, et al. CareTrack: assessing the appropriateness of health care delivery in Australia. Medical Journal of Australia 2012;**197**(10):549.
7 Macnaughton J. Anecdote in clinical practice. In: Greenhalgh T, Hurwitz B, editors. *Narrative basedmedicine: dialogue and discourse in clinical practice*. London: BMJ Publications; 1998.
8 Greenhalgh T. Narrative based medicine: narrative based medicine in an evidence based world. BMJ: British Medical Journal 1999;**318**(7179):323.
9 Greenhalgh T. Intuition and evidence – uneasy bedfellows? The British Journal of General Practice 2002;**52**(478):395.
10 Mori R, Lakhanpaul M, Verrier-Jones K. Guidelines: diagnosis and management of urinary tract infection in children: summary of NICE guidance. BMJ: British Medical Journal 2007;**335**(7616):395.

11 Mulrow CD. Rationale for systematic reviews. BMJ: British Medical Journal 1994;**309**(6954):597.
12 van Helmont JA. *Oriatrike, or physick refined: the common errors therein refuted and the whole art reformed and rectified*. London: Lodowick-Loyd, 1662.
13 Swinglehurst DA. Information needs of United Kingdom primary care clinicians. Health Information & Libraries Journal 2005;**22**(3):196-204.
14 Smith R. Where is the wisdom…? BMJ: British Medical Journal 1991;**303**(6806):798.
15 Straus SE, Richardson WS, Glasziou P, et al. *Evidence-based medicine: how to practice and teach EBM* (Fourth Edition). Edinburgh: Churchill Livingstone; 2010.

Capítulo 2 **Búsquedas en la literatura**

La evidencia se acumula cada vez más rápido y mantenerse actualizado es esencial para ofrecer una asistencia de calidad a los pacientes.

Los estudios y revisiones de estudios sobre la conducta de búsqueda de información de los médicos confirman que los manuales y los contactos personales siguen siendo las fuentes preferidas de información clínica, seguidos de los artículos de revistas[1]. El uso de internet como fuente de información ha aumentado de forma espectacular en los últimos años, sobre todo a través de PubMed/Medline, pero la sofisticación de la búsqueda y la eficacia a la hora de encontrar las respuestas no han aumentado de forma sustancial. De hecho, cualquier bibliotecario médico podría contar que ciertas preguntas clínicas importantes se abordan mediante búsquedas no sistemáticas en Google. Aunque la necesidad que los profesionales sanitarios tienen de obtener información de la máxima calidad nunca ha sido mayor, existen numerosos obstáculos: falta de tiempo, déficit de medios, carencia de habilidades de búsqueda, falta de motivación y (tal vez lo peor de todo) sobrecarga de información[2].

La literatura médica se ha convertido en una jungla mucho más intrincada de lo que era cuando se publicó la primera edición de este libro en 1996. El volumen y la complejidad de la literatura publicada se han disparado: sólo Medline tiene más de 20 millones de referencias. Aunque Medline es el buque insignia de las bases de datos que recogen artículos de revistas de ciencias de la salud, es un recurso muy conservador y recopila con lentitud las revistas nuevas o las publicadas fuera de EE.UU., por lo que hay muchos miles de artículos de gran calidad que pueden estar disponibles a través de otras bases de datos, pero que no se incluyen en los 20 millones de Medline. La proliferación de bases de datos hace que la selva de información sea mucho más confusa, sobre todo porque cada base de datos abarca su propia gama de revistas y cada una tiene sus propios protocolos de búsqueda particulares. ¿Cómo se puede hacer frente a esto?

No hay que desesperarse: en la última década, la «jungla» de información se ha domesticado gracias a las autopistas de información y los sistemas de transmisión de alta velocidad. Los conocimientos sobre el modo de acceder a estas maravillas para la navegación ahorrarán tiempo y mejorarán la capacidad de encontrar la mejor evidencia. El objetivo de este capítulo no es enseñar al lector a convertirse en un buscador experto sino ayudarle a reconocer los tipos de recursos que están disponibles, elegir inteligentemente entre ellos y aprovecharlos directamente.

¿Qué estamos buscando?

La literatura médica (y, más ampliamente, la de las ciencias de la salud) se puede consultar por tres motivos generales:

- De manera informal, casi con fines recreativos, navegando para mantenernos actualizados y satisfacer nuestra curiosidad intrínseca.
- De forma dirigida, en busca de respuestas, que pueden estar relacionadas con las preguntas que han surgido en la clínica o que se derivan de los pacientes y sus preguntas.
- Evaluación de la literatura existente, quizás antes de embarcarse en un proyecto de investigación.
- Cada uno de estos enfoques implica buscar de una manera muy diferente.

La *navegación* conlleva cierto grado de azar. Antes de la revolución digital, bastaba con abrir nuestra revista favorita y hojear sus páginas para ver qué se había escrito. Si contábamos con alguna herramienta para ayudarnos a discriminar la calidad de los artículos que encontrábamos, mucho mejor. En la actualidad, es posible utilizar estas herramientas para ayudarnos a navegar en este mar de publicaciones. Es posible navegar por las revistas electrónicas con la misma facilidad que por las revistas en papel; hoy en día se pueden utilizar servicios de alertas para recibir avisos de cuándo se ha publicado un nuevo número e incluso para indicarnos si ese número incluye artículos que se ajusten a nuestro perfil de interés. Ahora es posible recibir información en formato RSS (acrónimo en inglés de «resumen óptimo del sitio») de artículos de revistas específicas o sobre temas específicos en nuestro correo electrónico, teléfono móvil o blogs personales, y podemos participar en intercambios de Twitter relacionados con los artículos recién publicados. Casi todas las revistas tienen vínculos desde su página de inicio que permiten acceder al menos a uno de estos servicios de redes sociales. Estas tecnologías están cambiando continuamente. Aquellos de nosotros que nos hemos enfrentado a aluviones de publicaciones nuevas, fotocopias y revistas con la intención de leerlas estaremos encantados de saber que podemos disfrutar de ese mismo caos en formato electrónico. En eso consiste navegar en busca de fortuna y es algo maravilloso que nunca desaparecerá, con independencia del formato en que se publique nuestra literatura.

Buscar respuestas implica un enfoque mucho más dirigido, pues se trata de buscar una respuesta que podamos aplicar directamente a la asistencia de un paciente de un modo fiable. Cuando se encuentra esa información de confianza, se puede dejar de buscar. No es necesario escudriñar todos y cada uno de los estudios que puedan haberse publicado sobre el tema. Este tipo de consulta cada vez se basa más en las nuevas fuentes de información sintetizada, cuyo objetivo es apoyar la asistencia basada en la evidencia y la transferencia de resultados de la investigación a la práctica. Esto se describe con más detalle en este capítulo.

La evaluación de la literatura (es decir, la preparación de una revisión detallada, extensa y reflexiva de la literatura, por ejemplo, al escribir un ensayo para un trabajo o un artículo para su publicación) implica un proceso completamente

diferente. El propósito en este caso no es tanto influir en la asistencia del paciente directamente sino identificar el conjunto existente de investigaciones que han abordado un problema y aclarar las lagunas existentes en los conocimientos que requieran más investigaciones. Para este tipo de búsqueda, es fundamental contar con amplios conocimientos sobre los recursos de información y con habilidades a la hora de buscarlos. Una simple búsqueda en PubMed no será suficiente. Habrá que buscar de forma sistemática en múltiples bases de datos pertinentes y es necesario utilizar el encadenamiento de citas (v. más adelante) para asegurarse de no dejar ni una piedra sin remover. Si éste es nuestro objetivo, debemos consultar con un profesional de la información (bibliotecario de ciencias de la salud, profesional de la información clínica, etc.).

Jerarquía de los niveles de evidencia

El término *nivel de evidencia* se refiere al grado de confianza, basada en el diseño del estudio, que puede otorgarse a la información. Tradicionalmente, y teniendo en cuenta el tipo más frecuente de pregunta (en relación al tratamiento), los niveles de evidencia se representan como una pirámide en la que las revisiones sistemáticas se sitúan con grandiosidad en la parte superior, seguidas de los ensayos controlados y aleatorizados bien diseñados, después los estudios observacionales, como los estudios de cohortes o de casos y controles, para finalizar con los estudios de casos, estudios de laboratorio y la «opinión de expertos» en algún lugar cerca de la base (fig. 2.1). Esta jerarquía tradicional se describe con más detalle en la sección «Jerarquía tradicional de la evidencia».

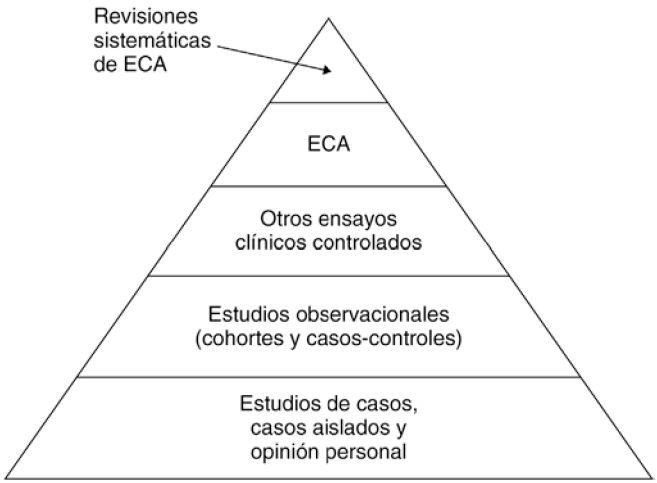

Figura 2.1 Jerarquía simple de la evidencia para evaluar la calidad del diseño de los ensayos en los estudios sobre tratamientos. ECA, ensayos controlados aleatorizados.

Mis colegas bibliotecarios, que suelen estar interesados en la evidencia resumida y en los recursos técnicos de apoyo para la toma de decisiones, me han hablado de una pirámide rival, en la que los sistemas informatizados de apoyo para la toma de decisiones se sitúan en la parte superior, por encima de las guías de práctica clínica basadas en la evidencia, seguidas de las sinopsis de revisiones sistemáticas, que están por encima de las revisiones no sistemáticas, etc.[3].

Tanto si pensamos en términos de la primera pirámide de evidencia (tradicional) como de la segunda (más actual), el mensaje está claro: toda la evidencia, toda la información, no es necesariamente equivalente. Debemos estar ojo avizor para analizar la credibilidad de cualquier información que nos encontremos, dondequiera que la encontremos.

Fuentes sintetizadas: sistemas, resúmenes y síntesis

Los recursos de información sintetizados a partir de los estudios primarios representan un nivel de evidencia muy alto. Estos recursos se han creado para ayudar a transformar la investigación en práctica y documentar la toma de decisiones del médico y del paciente. Este tipo de evidencia es relativamente nuevo (al menos, en comparación con los estudios de investigación primarios tradicionales, que nos han acompañado durante siglos), pero es de esperar que su uso crezca considerablemente a medida que se conozcan mejor.

Las *revisiones sistemáticas* son quizá la fuente sintetizada más antigua y más conocida. Su inicio se remonta a la década de 1980, auspiciadas por Archie Cochrane, quien se lamentaba de la multiplicidad de ensayos clínicos individuales cuya información no lograba ofrecer mensajes claros aplicables en la práctica. Los primeros intentos de realizar búsquedas exhaustivas de ensayos clínicos sobre un tema y agrupar sus resultados estadísticamente se convirtieron en la Biblioteca Cochrane (Cochrane Library) a mediados de la década de 1990; las Revisiones Cochrane pasaron a ser el patrón oro de las revisiones sistemáticas y la Colaboración Cochrane se convirtió en el estímulo principal para desarrollar y mejorar la metodología de revisión[4].

Las revisiones sistemáticas presentan muchas ventajas, pero hay que tener ciertas precauciones. Por el lado positivo, las revisiones sistemáticas son relativamente fáciles de interpretar. La selección sistemática y la evaluación de los estudios primarios según un protocolo aprobado significan que el sesgo se minimiza. Los estudios más pequeños, que son demasiado frecuentes en algunas áreas temáticas, pueden mostrar una tendencia hacia un impacto positivo, pero carecen de significación estadística. Sin embargo, cuando los datos de varios estudios pequeños se suman matemáticamente en un proceso denominado *metaanálisis,* los datos combinados pueden producir un resultado estadísticamente significativo (v. sección «Metaanálisis para los no estadísticos»). Las revisiones sistemáticas pueden ayudar a solucionar los resultados contradictorios entre los diferentes estudios sobre la misma cuestión. Si la revisión sistemática se ha realizado correctamente, es probable que los resultados sean sólidos y generalizables. Por

el lado negativo, las revisiones sistemáticas pueden replicar y ampliar los fallos de los estudios primarios (p. ej., si todos los estudios primarios evaluaron un fármaco a dosis subterapéutica, la conclusión global [engañosa] puede ser que el fármaco no tiene «ningún efecto»). Las Revisiones Cochrane pueden ser una lectura abrumadora, pero el siguiente consejo hará que todo parezca más fácil. El grueso de una Revisión Cochrane consiste en una discusión metodológica cuya esencia se presenta en el «Resumen en términos sencillos» *(Plain Language Summary),* que siempre se encuentra inmediatamente después del resumen. Como alternativa, se puede obtener un resumen rápido y preciso observando los gráficos, sobre todo el denominado *gráfico* o *diagrama de bosque (forest plot),* que muestra gráficamente los resultados de cada uno de los estudios primarios, junto con el resultado combinado (metaanálisis). En el capítulo 9 se explican las revisiones sistemáticas con más detalle.

Las Revisiones Cochrane sólo se publican en formato electrónico, pero la literatura clínica recoge otras revisiones sistemáticas. La forma más fácil de acceder a ellas es a través de la Biblioteca Cochrane, que publica las Revisiones Cochrane, la base de datos DARE (acrónimo inglés de «Base de Datos de Resúmenes de Revisiones de Efectos», recogidos en la Biblioteca Cochrane como «Otras revisiones» *[Other Reviews]*), y una base de datos de evaluaciones de tecnología sanitaria *(Health Technology Assessments, HTA).* La base de datos DARE no sólo proporciona una bibliografía de revisiones sistemáticas sino también una valoración crítica de la mayoría de las revisiones incluidas, lo que la convierte en un «fuente preevaluada» para las revisiones sistemáticas. Las HTA son esencialmente revisiones sistemáticas, pero van un paso más allá, para tener en cuenta las implicaciones económicas y políticas de los fármacos, tecnologías y sistemas de salud. En todas estas fuentes se pueden realizar búsquedas de manera relativamente simple y simultánea a través de la Biblioteca Cochrane.

Antiguamente, las Revisiones Cochrane se centraban principalmente en cuestiones relativas al tratamiento (v. cap. 6) o a la prevención, pero desde el año 2008 se han llevado a cabo esfuerzos considerables para la elaboración de revisiones sistemáticas de pruebas diagnósticas (v. cap. 8).

Los *recursos de punto de asistencia (Point of Care Resources)* son como libros de texto electrónicos o manuales clínicos detallados, pero están explícitamente basados en la evidencia, se actualizan de forma continua y están diseñados para ser fáciles de usar. Es posible que sean los libros de texto del futuro. Tres de los más conocidos son *Clinical Evidence, DynaMed* y *American College of Physicians Physicians' Information and Education Resource (ACP PIER).* Todos ellos aspiran a estar firmemente basados en la evidencia, a ser revisados por expertos y con regularidad, así como a incluir en sus recomendaciones enlaces con la investigación primaria.

- *Clinical Evidence* (http://clinicalevidence.bmj.com) es un recurso británico que se basa en revisiones sistemáticas para proporcionar información muy rápida, en especial sobre el valor comparativo de las pruebas e intervenciones. Las revisiones se organizan en secciones, como salud infantil o trastornos cutáneos

(*Child Health, Skin Disorders*); también se pueden buscar por palabras clave (p. ej., «asma») o por una lista completa de revisiones. La página inicial de un capítulo recoge las preguntas sobre la eficacia de diversas intervenciones y muestra señalizadores con medallones de color dorado, blanco o rojo para indicar si la evidencia existente es positiva, equívoca o negativa.

- *DynaMed* (http://www.ebscohost.com/dynamed/), es un recurso estadounidense organizado como un manual con capítulos que abarcan una amplia variedad de trastornos, pero con resúmenes de la investigación clínica, los niveles de evidencia y enlaces con los artículos primarios. Abarca las causas y riesgos, complicaciones y las afecciones asociadas (incluido el diagnóstico diferencial), lo que se debe buscar en la anamnesis y la exploración física, las pruebas diagnósticas que se deben hacer, el pronóstico, el tratamiento, la prevención y el cribado, así como enlaces a folletos de información a los pacientes. Es muy fácil realizar una búsqueda sobre un trastorno específico: los resultados incluyen enlaces a otros capítulos sobre trastornos similares. Se trata de un recurso de pago, aunque puede ser gratuito para quienes se ofrecen a escribir un capítulo.
- El *ACP PIER* (*American College of Physicians Physicians' Information and Education Resource:* http://pier.acponline.org) es otra fuente estadounidense. Utiliza el formato estándar de recomendación amplia, recomendación específica, fundamentos y evidencia. El ACP PIER abarca la prevención, cribado, diagnóstico, consulta, hospitalización, tratamientos farmacológicos y no farmacológicos y el seguimiento. Ofrece los enlaces a la literatura primaria y la pestaña «Información para el paciente» proporciona enlaces a páginas web que serán útiles y fidedignas para los pacientes.

Tanto el PIER como DynaMed tienen aplicaciones que facilitan su empleo en agendas electrónicas (PDA) y otros dispositivos portátiles, lo que mejora su usabilidad a la hora de prestar asistencia a la cabecera del paciente.

Continuamente están surgiendo nuevos recursos de punto de asistencia, por lo que escoger cuál se usa depende en gran medida de la preferencia personal. Los tres que se han descrito se han elegido porque están revisados por expertos, se actualizan regularmente y tienen enlaces directos a la evidencia primaria.

Las *guías de práctica clínica,* que se describen con detalle en el capítulo 10, son «recomendaciones elaboradas de manera sistemática para ayudar en la toma de decisiones del médico y del paciente sobre la atención médica apropiada en circunstancias clínicas específicas»[5]. En una guía práctica clínica de buena calidad, la evidencia científica se reúne de manera sistemática, los autores que elaboran la guía son representantes de todas las disciplinas pertinentes, incluidos los pacientes, y las recomendaciones están vinculadas explícitamente con la evidencia de la que se derivan[6]. Las guías de práctica clínica son una forma resumida de evidencia y ocupan un lugar muy alto en la jerarquía de los recursos preevaluados, pero el propósito inicial se debería tener siempre en cuenta: las guías para diferentes contextos y distintos propósitos se pueden basar en la misma evidencia, pero proporcionan diferentes recomendaciones.

Las guías son fácilmente accesibles y se encuentran en diversas fuentes, como las que se recogen a continuación.

- *National Guideline Clearinghouse* (http://www.guideline.gov/). Es una iniciativa de la Agency for Healthcare Research and Quality (AHRQ), del Department of Health and Human Services estadounidense. Aunque es una base de datos con financiación del gobierno de EE.UU., el National Geographic Channel (NGC) cuenta con contenido internacional. Una ventaja de este recurso es que las diferentes guías que pretenden abarcar el mismo tema pueden compararse directamente en todos los aspectos, desde los niveles de evidencia hasta las recomendaciones. Todas las guías deben estar actualizadas y revisarse cada 5 años.
- *National Institute for Health and Care Excellence* (*NICE*, http://www.nice.org.uk/). Es un organismo financiado por el gobierno de Reino Unido, responsable de la elaboración de guías basadas en la evidencia y de otros resúmenes de evidencia para apoyar la política sanitaria británica. Los NICE Clinical Knowledge Summaries (http://cks.nice.org.uk) están diseñados especialmente para los profesionales de atención primaria.

Una forma sencilla y popular de buscar guías prácticas es a través de TRIP (Turning Research into Practice, http://www.tripdatabase.com), un motor de búsqueda federada que se describe más adelante. Para ello se debe consultar la lista desplegable que se muestra a la derecha de la pantalla después de una búsqueda simple: en ella aparece el encabezado «*Guidelines*» (guías), con subencabezados para Australia y Nueva Zelanda, Canadá, Reino Unido, EE.UU. y otros países, así como una cifra que indica el número de guías encontradas sobre ese tema. Las fuentes NICE y National Guideline Clearinghouse se incluyen entre las guías buscadas.

Fuentes preevaluadas: sinopsis de revisiones sistemáticas y estudios primarios

Si nuestro tema de interés es más limitado que los que se recogen en las fuentes sintetizadas o resumidas que se han descrito, o si simplemente se realiza una búsqueda para mantenerse al día con la literatura, se puede consultar una de las fuentes preevaluadas como medio de navegación a través de los millones de artículos existentes en nuestra jungla de información. El formato más común es el compendio de artículos de investigación clínica extraídos de revistas principales y que se considera una fuente de información importante para la atención del paciente: *Evidence-Based Medicine, ACP Journal Club, Evidence-Based Mental Health* o *POEMS (Patient-Oriented Evidence that Matters)*. Algunos son gratuitos y otros están disponibles a través de instituciones, afiliaciones o suscripción privada. Todos ellos tienen un formato que incluye un resumen estructurado y una breve valoración crítica del contenido del artículo. Los estudios incluidos pueden ser estudios individuales o revisiones sistemáticas. Cada uno se considera una fuente preevaluada y, aparte de la evaluación crítica, la simple inclusión tiene implicaciones para la calidad percibida y la importancia del artículo original.

Todas estas fuentes pueden considerarse pequeñas bases de datos de estudios seleccionados, donde pueden realizarse búsquedas por palabras clave. Otros servicios seleccionados de artículos de revistas, como Evidence Updates, proporcionan resúmenes más una indicación del nivel de interés que cada artículo podría tener para disciplinas específicas.

DARE se ha mencionado como una fuente preevaluada para las revisiones sistemáticas distintas de las Revisiones Cochrane ya que proporciona un resumen ampliado y una breve evaluación crítica de la mayoría de las revisiones sistemáticas incluidas en su base de datos.

Otra fuente que se considera preevaluada, aunque no contiene evaluaciones, es el *Central Register of Controlled Trials,* que también forma parte de la Biblioteca Cochrane. «Central» es una base de datos de los estudios incluidos en las Revisiones Cochrane, así como de los nuevos estudios sobre temas similares, mantenidos por los distintos Grupos de Revisión Cochrane. En la Biblioteca Cochrane es posible realizar búsquedas de forma simultánea en DARE, Central, la Base de Datos Cochrane de Revisiones Sistemáticas, la base de datos HTA y la NHS Economic Evaluation Database (que también incluye resúmenes de los estudios evaluados de forma crítica).

Recursos especializados

Las fuentes de información especializadas, organizadas (como su nombre indica) para ayudar al médico especialista en un campo en particular, también suelen ser útiles para los médicos generales, enfermeras especializadas y médicos de atención primaria. La mayoría de las asociaciones profesionales mantienen páginas web excelentes con guías de práctica clínica, enlaces a revistas y otros recursos de información útiles; la mayoría requiere ser miembro de la asociación para tener acceso a los materiales educativos y prácticos. Tres ejemplos notables que están disponibles pagando una cuota son Global Infectious Diseases and Epidemiology Network (GIDEON), Psychiatry Online y CardioSource.

- *GIDEON* (*Global Infectious Diseases and Epidemiology Network*, http://www.gideononline.com/) es un programa basado en la evidencia que ayuda en el diagnóstico y tratamiento de las enfermedades contagiosas. Además, GIDEON realiza un seguimiento de la incidencia y prevalencia de enfermedades en todo el mundo e incluye el espectro cubierto por los antibióticos.
- *Psychiatry Online* (http://www.psychiatryonline.com/) es un compendio de manuales fundamentales (incluida la quinta edición del *Manual diagnóstico y estadístico de los trastornos mentales* [DSM-5]), revistas de psiquiatría y guías de práctica clínica de la American Psychiatric Association, elaboradas por la American Psychiatric Press.
- *CardioSource* (http://www.cardiosource.com) está elaborada por el American College of Cardiology. Incluye guías, enlaces a revistas y libros, «colecciones clínicas» de artículos y materiales educativos sobre temas como el control del colesterol y la fibrilación auricular, y un registro excelente de ensayos clínicos

para todos los ensayos referentes a las enfermedades cardiovasculares, tanto activos como finalizados.

Éstos son sólo tres ejemplos. Con independencia de cuál sea nuestra especialidad (o tema especializado), por lo general existirá un recurso semejante mantenido por una sociedad profesional. Es posible preguntar a un bibliotecario o a un profesional de la información clínica para que nos ayude a encontrar el que sea relevante.

Estudios primarios: desentrañando la selva

Ya sea por hábito o por falta de familiaridad con todas las fuentes útiles sintetizadas, resumidas o preevaluadas descritas en los párrafos previos, la mayoría de los profesionales sanitarios prefieren una búsqueda básica de Medline/PubMed para satisfacer sus necesidades de información clínica. Algunos simplemente prefieren evaluar la literatura primaria por sí mismos, sin valoraciones críticas resumidas ni su inclusión en grandes revisiones sobre el tratamiento de enfermedades. Mi consejo sigue siendo que se consulten las fuentes secundarias descritas en las secciones «Jerarquía de los niveles de evidencia», «Fuentes sintetizadas: sistemas, resúmenes y síntesis» y «Fuentes preevaluadas: sinopsis de revisiones sistemáticas y estudios primarios». No obstante, si el lector prefiere ir directamente a los estudios primarios, esta sección le será de utilidad.

Las fuentes primarias se pueden encontrar de varias formas. Pueden consultarse las listas de referencias e hipervínculos de las fuentes secundarias descritas. También se pueden identificar a partir de las revistas, por ejemplo, a través de canales RSS, servicios de tabla de contenidos o servicios de información temática más específicos. También se pueden realizar búsquedas en bases de datos como PubMed/Medline, EMBASE, PASCAL, Cochrane Library, CINAHL (Cumulated Index of Nursing and Allied Health Literature), Biosis Previews, Web of Science, Scopus, Google o Google Scholar. A continuación, se describirán todas estas fuentes de forma individual.

PubMed es el recurso de internet que más visita la mayoría de los médicos y profesionales sanitarios de todo el mundo, probablemente debido a que es gratuito y muy conocido. La mayoría de la gente opta por una búsqueda básica en PubMed, utilizando dos o tres palabras como texto de búsqueda en el mejor de los casos, lo que suele dar como resultado que aparezcan cientos o miles de referencias, de las cuales sólo se consultan las dos primeras pantallas. Es evidente que ésta no es la forma más eficaz de buscar, pero constituye la realidad de las búsquedas que *realiza* la mayoría de las personas[7]. Curiosamente, basta con añadir uno o dos términos de búsqueda adicionales para que la eficacia de una búsqueda básica en PubMed mejore sustancialmente[7].

Se pueden utilizar herramientas sencillas que forman parte del motor de búsqueda de Medline para ayudar a centrar una búsqueda y obtener mejores resultados de una búsqueda básica, pero los estudiantes de medicina o los médicos pocas veces los utilizan. Una de estas herramientas es la función *limit* (limitar),

que permite establecer restricciones a temas tan genéricos como sexo, grupo de edad o diseño del estudio, al idioma o a las revistas clínicas principales. La función de búsqueda avanzada de PubMed incorpora estos límites en una misma página de búsqueda. La próxima vez que navegue por la página web de PubMed con tiempo de sobra, practique con estas herramientas y vea lo fácil que puede resultar afinar su búsqueda.

Clinical queries (consultas clínicas) es una opción que aparece en el panel de la izquierda de la pantalla básica de PubMed o en la parte inferior de la pantalla de búsqueda avanzada. Esta opción superpone en la búsqueda un filtro basado en diseños de estudio óptimos para lograr la mejor evidencia, dependiendo del ámbito de la pregunta y del grado en que se desea centrar dicha pregunta. Por ejemplo, si queremos buscar un estudio sobre el tratamiento de la hipercolesterolemia (en inglés, *hypercholesterolaemia*), la *clinical query* para *therapy/narrow and specific* tendría que plantearse del siguiente modo «(hypercholesterolaemia) AND (randomised controlled trial [Publication Type] OR (randomised [Title/Abstract] AND controlled [Title/Abstract] AND trial [Title/Abstract]))». En este caso, la búsqueda podría necesitar más límites o tal vez la adición de un segundo término, como un fármaco específico, ya que el resultado incluye más de 2.000 publicaciones.

La función *citation chaining* (encadenamiento de citas, también denominada *citation tracking*) proporciona otro medio de seguir un tema. Por ejemplo, puede que después de nuestra búsqueda sobre la hipercolesterolemia queramos seguir un estudio de investigación clásico primario, como el West of Scotland Coronary Prevention Study, publicado inicialmente en la década de 1990[8]. En nuestra búsqueda en PubMed encontramos un estudio en el New England Journal of Medicine de 2007 que describe un seguimiento de 20 años[9], pero nos surge la duda de si se ha publicado algo más al respecto. Las bases de datos Web of Science, que engloban el Science Citation Index, el Social Sciences Citation Index y el Arts and Humanities Citation Index en línea, ofrecen una función de búsqueda de referencias citadas. Si introducimos el nombre del autor (en este caso I. Ford) y el año de publicación (2007), podemos hacer un seguimiento del artículo específico y encontrar varias decenas de artículos publicados desde entonces donde se le cita en sus listas de referencias. La búsqueda de citas puede dar una indicación aproximada de la importancia relativa de un estudio, basada en el número de veces que ha sido citado (sin olvidar que a veces se cita un artículo criticando su mala calidad). Una forma muy sencilla (pero algo menos precisa) de encadenamiento de citas es utilizar Google Académico (Google Scholar en su versión en inglés): basta con introducir el título del artículo en este motor de búsqueda y, cuando se ha encontrado, seleccionar «citas» *(citations)*.

Google Académico es un navegador de internet muy amplio, cada vez más popular y muy práctico, ya que es accesible desde la barra de herramientas de Google. Para un tema poco conocido, Google Académico puede ser un excelente recurso al que recurrir, ya que identificará tanto los artículos que se recogen en PubMed como los que no. Por desgracia, carece de filtros de calidad (como *clinical*

queries) y de límites (como el sexo o la edad), por lo que una búsqueda sobre un tema muy investigado tenderá a devolver una extensa lista de resultados en la que no habrá más remedio que buscar de forma manual.

Sistema de ventanilla única: motores de búsqueda federada

Tal vez la alternativa más sencilla y más eficaz para la mayoría de los médicos que buscan información para la atención al paciente sea un motor de búsqueda federada como TRIP, http://www.tripdatabase.com/, que busca en múltiples fuentes de forma simultánea y tiene la ventaja de ser gratuito.

TRIP tiene un motor de búsqueda muy primitivo, pero busca en fuentes sintetizadas (revisiones sistemáticas, incluidas las Revisiones Cochrane), fuentes resumidas (como las guías prácticas de Norteamérica, Europa, Australia/Nueva Zelanda y otras regiones, así como libros de texto electrónicos) y en fuentes preevaluadas (como las revistas *Evidence-Based Medicine y Evidence-Based Mental Health*), además de buscar en todos los ámbitos de *Clinical Queries* de PubMed simultáneamente. Además, las búsquedas pueden limitarse por disciplina, como pediatría o cirugía, lo que ayuda tanto a centrar la búsqueda como a eliminar resultados claramente irrelevantes y también tiene en cuenta la tendencia de los especialistas médicos a preferir (correcta o incorrectamente) la literatura de sus propias revistas. Dado que la mayoría de los médicos prefieren búsquedas muy simples, una búsqueda TRIP puede ser la mejor para rentabilizar al máximo el tiempo y las energías disponibles para dicha búsqueda.

Fuentes de ayuda y preguntas a conocidos

Si un bibliotecario se fracturara la muñeca, no dudaría en ir al médico. Del mismo modo, un profesional sanitario no tiene que enfrentarse a la literatura en soledad. Los bibliotecarios del ámbito de la salud están fácilmente accesibles en las universidades, hospitales, departamentos y organismos gubernamentales, y sociedades profesionales. Estos profesionales conocen las bases de datos disponibles, son conscientes de las complejidades de las búsquedas, conocen la literatura (incluso los documentos gubernamentales complejos y los grupos de datos poco claros) y suelen saber lo suficiente sobre el tema para tener una idea de lo que estamos buscando y los niveles de evidencia que es probable encontrar. Si un bibliotecario no puede encontrar una respuesta, puede consultar con sus compañeros a nivel local, nacional e internacional. Los bibliotecarios del siglo XXI están muy bien conectados en red.

Preguntar a personas conocidas tiene sus ventajas. Los expertos en un campo suelen conocer investigaciones no publicadas o informes encargados por el gobierno u otras agencias, sobre todo la literatura «gris» o «fugitiva» muy difícil de encontrar que no está indexada en ninguna fuente. La organización internacional de intercambio de información CHAIN (Contact, Help, Advice and Information

Network, http://chain.ulcc.ac.uk/chain) constituye una red en línea muy útil para las personas que trabajan en asistencia sanitaria y social que deseen compartir información. Es posible unirse a CHAIN de forma gratuita y, una vez convertidos en miembros, se pueden plantear preguntas y dirigirlas a un grupo específico de especialistas.

En un campo tan abrumador y complejo como la información de salud, preguntar a los colegas y a personas de confianza ha sido siempre una de las fuentes preferidas de información. En los primeros días de la medicina basada en la evidencia (MBE), preguntar a los conocidos se consideraba poco sistemático y «sesgado». Sigue siendo cierto que preguntar a los conocidos no es suficiente a la hora de buscar evidencia, pero dada la capacidad de los expertos a la hora de identificar literatura poco conocida, tal vez ninguna búsqueda podría considerarse completa sin recurrir a este método.

Tutoriales en línea para una búsqueda eficaz

Muchas universidades y otras instituciones educativas proporcionan actualmente tutoriales de autoaprendizaje a los que se puede acceder a través del ordenador, ya sea en una intranet (sólo para los miembros de la universidad) o en internet (accesible a todo el mundo). A continuación, se presentan algunos que encontré al revisar este capítulo para la quinta edición. El lector debe tener en cuenta que, al igual que sucede con todas las fuentes basadas en internet, algunos sitios pueden cambiar de dirección o cerrar, por lo que pido disculpas de antemano si alguno de los enlaces no lleva a buen puerto:

- «*Finding the Evidence*» del Centro de Medicina Basada en la Evidencia de la Universidad de Oxford. Ofrece varios consejos breves sobre las búsquedas en bases de datos clave, pero no enseña mucho sobre el modo de realizarlas. Tal vez esté más indicado para quienes ya han realizado un curso y quieran refrescar su memoria. http://www.cebm.net/index.aspx?o = 1038.
- «*PubMed – Searching Medical Literature*» de la biblioteca de la Georgia State University. Como el título indica, este tutorial se limita a PubMed, pero ofrece algunos trucos avanzados, como la forma de personalizar la interfaz de PubMed para adaptarla a las necesidades personales. http://research.library.gsu.edu/pubmed.
- «*PubMed Tutorial*» del propio PubMed. Ofrece una visión general de lo que PubMed hace y no hace, así como algunos ejercicios para acostumbrarse a utilizarlo. http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/.

Hay muchos otros tutoriales similares que pueden consultarse gratuitamente en internet, pero pocos de ellos van mucho más allá de la búsqueda de estudios primarios y revisiones sistemáticas en PubMed y en la Biblioteca Cochrane. Espero que cuando se publique la próxima edición de este libro, alguien haya corregido este sesgo y se hayan elaborado tutoriales sobre cómo acceder a toda la gama de resúmenes, síntesis y fuentes preevaluadas que se han descrito en las secciones anteriores.

Bibliografía

1 Davies K. The information seeking behaviour of doctors: a review of the evidence. Health Information & Libraries Journal 2007;**24**(2):78-94.
2 Fourie I. Learning from research on the information behaviour of healthcare professionals: a review of the literature 2004-2008 with a focus on emotion. Health Information & Libraries Journal 2009;**26**(3):171-86.
3 DiCenso A, Bayley L, Haynes R, ACP Journal Club. Editorial: accessing preappraised evidence: fine-tuning the 5S model into a 6S model. Annals of Internal Medicine 2009;**151**(6): JC3.
4 Levin A. The Cochrane collaboration. Annals of Internal Medicine 2001;**135**(4):309-12.
5 Field MJ, Lohr KN. *Clinical practice guidelines: directions for a new program*. Washington, DC: National Academy Press, 1990.
6 Grimshaw J, Freemantle N, Wallace S, et al. Developing and implementing clinical practice guidelines. Quality in Health Care 1995;**4**(1):55.
7 Hoogendam A, Stalenhoef AF, de Vries Robbé PF, et al. Answers to questions posed during daily patient care aremore likely to be answered by UpToDate than PubMed. Journal of Medical Internet Research 2008;**10**(4):e29.
8 Shepherd J, Cobbe SM, Ford I, et al. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. The New England Journal of Medicine 1995;**333**(20):1301-7 doi: 10.1056/nejm199511163332001.
9 Ford I, Murray H, Packard CJ, et al. Long-term follow-up of the West of Scotland Coronary Prevention Study. The New England Journal of Medicine 2007;**357**(15):1477-86 doi: 10.1056/NEJMoa065994.

Capítulo 3 Aprendiendo a orientarse: ¿de qué trata este artículo?

La ciencia de criticar los artículos

Por lo general, los estudiantes suelen sorprenderse cuando descubren que algunos (los puristas dirían que hasta el 99%) de los artículos publicados deberían acabar en la papelera y no se deberían utilizar para documentar la práctica. En 1979, el editor del British Medical Journal, el Dr. Stephen Lock, escribió: «pocas cosas son más desalentadoras para un editor médico que tener que rechazar un artículo basado en una buena idea, pero con defectos metodológicos irremediables». Quince años después, Altman seguía afirmando que sólo el 1% de la investigación médica no presentaba defectos metodológicos[1]; más recientemente, este mismo autor confirmó que incluso los artículos publicados en revistas de «calidad» suelen presentar defectos graves y fundamentales[2]. En el cuadro 3.1 se muestran los principales defectos por los que los artículos son rechazados (y que están presentes en cierto grado en muchos de los que terminan publicados).

La mayoría de los artículos que aparecen en las revistas médicas en la actualidad se presentan más o menos en el formato estándar de Introducción, Métodos, Resultados y Discusión (IMRD): Introducción (¿*por qué* los autores decidieron llevar a cabo esta investigación en particular?), Métodos (¿*cómo* lo hicieron y cómo eligieron analizar sus resultados?), Resultados (¿*qué* encontraron?) y Discusión (lo que los autores piensan que *significan* los resultados). Para decidir si un trabajo es digno de lectura, hay que basarse en el diseño de la sección de métodos y no en el interés de la hipótesis, la naturaleza o el impacto potencial de los resultados, o la especulación de la discusión.

En cambio, la mala ciencia es mala con independencia de si el estudio abordó un problema clínico importante, de si los resultados son «estadísticamente significativos» (v. sección «Probabilidad y confianza»), de si las cosas cambiaron en la dirección que nos habría gustado y de si los hallazgos prometen beneficios inconmensurables para los pacientes o un ahorro para el servicio de salud. En sentido estricto, *si se va a criticar seriamente un artículo, hay que hacerlo incluso antes de mirar los resultados.*

Cuadro 3.1 Razones frecuentes por las que los artículos son rechazados para su publicación

1. El estudio no abordó una cuestión científica importante (v. sección «Tres preguntas preliminares para orientarse»).
2. El estudio no era original, es decir, alguien había realizado antes el mismo estudio u otro similar (v. sección «¿Era el estudio original?»).
3. El estudio en realidad no sometía a prueba la hipótesis de los autores (v. sección «Tres preguntas preliminares para orientarse»).
4. Debería haberse utilizado un diseño de estudio diferente (v. sección «Ensayos controlados aleatorizados»).
5. Las dificultades prácticas (p. ej., al reclutar a los participantes) hicieron que los autores relajaran el protocolo original del estudio (v. sección «¿El diseño del estudio era acertado?»).
6. El tamaño de la muestra era demasiado pequeño (v. sección «¿Se abordaron las cuestiones estadísticas preliminares?»).
7. El estudio carecía de grupo control o dicho grupo era inadecuado (v. sección «¿Se evitó o se minimizó el sesgo sistemático?»).
8. El análisis estadístico fue incorrecto o inadecuado (v. cap. 5).
9. Los autores han extraído conclusiones injustificadas a partir de sus datos.
10. Existe un conflicto de intereses significativo (p. ej., uno de los autores, o un patrocinador, podría obtener beneficios económicos de la publicación del artículo y no se ha comprobado que se aplicasen suficientes garantías para proteger contra el sesgo).
11. El artículo está tan mal escrito que es incomprensible.

Es mucho más fácil encontrar defectos en el trabajo de otros autores que lograr que la investigación propia sea metodológicamente perfecta. Cuando enseño evaluación crítica, suele haber alguien en el grupo que encuentra muy descortés criticar los proyectos de investigación a los cuales científicos entregados han dedicado los mejores años de sus vidas. Desde un punto de vista más pragmático, puede haber buenas razones prácticas por las cuales los autores del estudio no han realizado un trabajo perfecto y ellos saben igual que nosotros que su trabajo tendría mayor validez científica si no hubiese surgido tal o cual dificultad (prevista o imprevista) durante la realización del estudio.

La mayoría de las revistas científicas de calidad envían los artículos a un revisor para que ofrezca sus comentarios sobre su validez científica, su originalidad y su importancia antes de decidir si se publican. Este proceso se denomina *revisión por pares* (o por expertos), y se ha escrito mucho al respecto[3]. En el cuadro 3.1 se resumen los defectos frecuentes señalados por estos revisores.

La evaluación de la calidad metodológica (evaluación crítica) se ha descrito en detalle en la serie ampliamente citada dirigida por Gordon Guyatt, «Users'

Guides to the Medical Literature» (el listado integro de listas de comprobación y los enlaces al texto completo gratuito de la mayoría de las listas pueden consultarse en JAMA Evidence http://www.cche.net/usersguides/main.asp). Muchos médicos consideran que las guías estructuradas elaboradas por estos autores sobre cómo leer artículos referentes al tratamiento, diagnóstico, cribado, pronóstico, causalidad, calidad asistencial, análisis económico, revisión sistemática, investigación cualitativa, etc. son las listas de comprobación definitivas para la evaluación crítica. En el apéndice 1 se enumeran algunas listas de comprobación más sencillas que yo he elaborado a partir de las Users' Guides y de otras fuentes citadas al final de este capítulo, junto con algunas ideas propias. Si el lector tiene experiencia a la hora de leer revistas, estas listas de comprobación no necesitarán muchas explicaciones. Pero si todavía se tienen dificultades al empezar cuando se lee un artículo médico, trate de hacerse las preguntas preliminares de la siguiente sección.

Tres preguntas preliminares para orientarse

Primera pregunta: ¿cuál fue la pregunta de investigación y por qué era necesario el estudio?

La frase introductoria de un artículo de investigación debería indicar, en pocas palabras, cuál es el trasfondo de la investigación. Por ejemplo, «la colocación de drenajes transtimpánicos es un procedimiento frecuente en los niños y se ha sugerido que no todas las operaciones son clínicamente necesarias». Esta afirmación debe seguirse de una breve revisión de la literatura publicada, por ejemplo, «el estudio prospectivo de Gupta y Brown sobre la colocación de drenajes transtimpánicos demostró que...». Es exasperantemente frecuente que los autores se olviden de contextualizar su investigación, pues el trasfondo del problema suele ser tan claro como el agua para ellos cuando llegan a la fase de escritura del artículo.

Salvo que ya se haya indicado en la introducción, la sección de métodos del artículo debe plantear claramente la pregunta de investigación y/o la hipótesis que los autores han decidido poner a prueba. Por ejemplo: «el objetivo de este estudio fue determinar si la cirugía ambulatoria de la hernia era más segura y más aceptable para los pacientes que el procedimiento estándar con ingreso hospitalario».

Es posible encontrar que la pregunta de investigación se ha omitido inadvertidamente o, lo que es más común, que la información esté inmersa en mitad de alguno de los párrafos. Si la hipótesis de investigación principal se plantea de forma negativa (que suele ser lo habitual), como «la adición de metformina al tratamiento con dosis máximas de sulfonilurea no mejorará el control de la diabetes tipo 2», se denomina hipótesis *nula*. Los autores de un estudio pocas veces *creen* realmente en su hipótesis nula cuando se embarcan en su investigación. Como seres humanos que son, por lo general pretenden demostrar una diferencia entre los dos grupos de su estudio, pero el modo en que los científicos lo hacen es afirmando: «*supongamos* que no hay diferencia y ahora

tratemos de refutar esa teoría». Según las enseñanzas de Popper, este método *hipotético-deductivo* de establecer hipótesis falsas que luego se procede a evaluar es la esencia misma del método científico[4].

Si no se ha descubierto cuál es la pregunta de investigación de los autores cuando ya se lleva leída la mitad de la sección de métodos, es posible encontrarla en el primer párrafo de la discusión. Sin embargo, hay que recordar que no todos los estudios de investigación (incluso los de calidad) están diseñados para probar una única hipótesis definitiva. Los estudios de investigación *cualitativos,* que son tan válidos y tan necesarios como los estudios cuantitativos más convencionales (siempre que estén diseñados y realizados adecuadamente), tienen como finalidad analizar las cuestiones particulares de manera amplia y abierta para esclarecer determinadas cuestiones, generar o modificar hipótesis y priorizar áreas de estudio. Este tipo de investigación se explica con más detalle en el capítulo 12. En la actualidad, incluso la investigación cuantitativa (que es el tema de la mayor parte del resto de este libro) se considera como algo más que la evaluación de hipótesis. Como se afirma en la sección «Probabilidad y confianza», es absolutamente preferible hablar en términos de evaluación de la *solidez* de la evidencia existente sobre un tema en particular que sobre probar o refutar las hipótesis.

Segunda pregunta: ¿cuál fue el diseño de la investigación?

En primer lugar, hay que decidir si el artículo describe un estudio primario o secundario. Los estudios primarios describen la investigación de primera mano, mientras que los estudios secundarios intentan resumir y extraer conclusiones a partir de los estudios primarios. Los estudios primarios (a veces denominados *estudios empíricos*) constituyen el elemento básico de la mayoría de las investigaciones publicadas en revistas médicas y, por lo general, pueden clasificarse en una de estas cuatro categorías:

- *Experimentos de laboratorio,* en los que se realiza una actuación en un animal o un voluntario en un entorno artificial y controlado.
- *Ensayos clínicos,* que son un tipo de experimento en el que se realiza una intervención, que puede ser sencilla (p. ej., un fármaco; v. cap. 6) o compleja (p. ej., un programa educativo; v. cap. 7) en un grupo de participantes (es decir, los pacientes incluidos en el ensayo) que después se siguen para ver lo que les pasa.
- *Estudios,* en los que se mide una variable en un grupo de participantes (pacientes, profesionales sanitarios u otra muestra de individuos). Los estudios mediante cuestionarios (cap. 13) miden las opiniones, actitudes y conductas autorreferidas de las personas.
- *Estudios de casos organizacionales,* en los que el investigador narra una historia que trata de captar la complejidad de un esfuerzo de cambio (p. ej., un intento de aplicar la evidencia; cap. 14).

Los tipos más frecuentes de ensayos clínicos y de estudios se describen en secciones posteriores de este capítulo. El lector debería entender la terminología utilizada en la descripción del diseño de los estudios (v. tabla 3.1).

Tabla 3.1 Términos usados para describir las características del diseño de los estudios de investigación clínica

Término	Significado
Comparación de grupos paralelos	Cada grupo recibe un tratamiento diferente, comenzando simultáneamente en ambos grupos. En este caso, los resultados se analizan comparando los grupos
Comparación emparejada	Los participantes que reciben tratamientos diferentes se emparejan para equilibrar las posibles variables de confusión, como la edad y el sexo. Los resultados se analizan en términos de diferencias entre los pares participantes
Comparación intraparticipante	Los participantes se evalúan antes y después de una intervención, y los resultados se analizan en términos de cambios en cada participante
Simple ciego	Los participantes ignoraban qué tratamiento recibían
Doble ciego	Los investigadores también ignoraban el tratamiento administrado
Diseño cruzado	Cada participante recibió tanto el tratamiento de la intervención como el de control (de forma aleatorizada), a menudo dejando un período de *reposo farmacológico* sin tratamiento
Controlado con placebo	Los participantes del grupo control reciben un placebo (píldora inactiva) que debe tener un aspecto y sabor idénticos a la píldora activa. También se pueden realizar intervenciones quirúrgicas de tipo placebo (simuladas) en los ensayos clínicos de cirugía
Diseño factorial	Estudio que permite evaluar los efectos (tanto de forma separada como combinada) de más de una variable independiente sobre un resultado específico (p. ej., se puede usar un diseño factorial 2×2 para evaluar los efectos del placebo, de la aspirina sola, de la estreptocinasa sola o de la aspirina + estreptocinasa en el infarto de miocardio agudo[5])

La investigación secundaria está constituida por:

- Revisiones, que se describen en el capítulo 9 y se pueden dividir en:
  (a) *Revisiones (no sistemáticas)*, que resumen los estudios primarios.
  (b) *Revisiones sistemáticas*, que efectúan dicho resumen usando un método riguroso, transparente y auditable (es decir, verificable).
  (c) *Metaanálisis*, que integran los datos numéricos de más de un estudio.
- *Guías clínicas*, que se describen en el capítulo 10. Extraen conclusiones de los estudios primarios sobre cuál debería ser el modo de actuación de los médicos.
- *Análisis de decisiones*, que no se detallan en este libro, pero se describen en otras obras[6]. Utilizan los resultados de los estudios primarios para generar árboles de probabilidad que pueden utilizar los profesionales sanitarios y los pacientes en la toma de decisiones sobre el manejo clínico.
- *Análisis económicos*, que se describen brevemente en el capítulo 12 y con más detalle en otras obras[7]. Utilizan los resultados de los estudios primarios para indicar si una línea de actuación determinada supone un uso adecuado de los recursos.

Tercera pregunta: ¿fue adecuado el diseño de investigación para la pregunta?

En las siguientes secciones se presentan ejemplos del tipo de preguntas que pueden responderse razonablemente mediante los diferentes tipos de estudio de investigación primario. Una pregunta que suele ser ineludible plantear es la siguiente: ¿un ensayo controlado aleatorizado (ECA) (v. sección «Ensayos controlados aleatorizados») era el mejor método de abordar esta pregunta de investigación particular y, si el estudio no era un ECA, debería haberlo sido? Antes de establecer ninguna conclusión, debe decidirse cuál es el ámbito general de la investigación que abarca el estudio (v. cuadro 3.2). Una vez hecho esto, hay que preguntarse si el diseño del estudio fue apropiado para esta pregunta.

Cuadro 3.2 Ámbitos generales de investigación

La mayoría de los estudios cuantitativos se refieren a uno o más de los siguientes ámbitos:

- *Tratamiento:* se evalúa la eficacia de los tratamientos farmacológicos, procedimientos quirúrgicos, métodos alternativos de prestación de servicios u otras intervenciones. El diseño de estudio preferido es el ensayo controlado aleatorizado (v. sección «Ensayos controlados aleatorizados» y los caps. 6 y 7).
- *Diagnóstico:* se intenta demostrar si una nueva prueba diagnóstica es válida (¿se puede confiar en ella?) y fiable (¿se obtendrían siempre los mismos resultados?). El diseño de estudio preferido es el estudio transversal (v. sección «Estudios transversales» y el cap. 8).
- *Cribado:* se trata de demostrar la utilidad de las pruebas que se pueden aplicar a grandes poblaciones y que detectan la enfermedad en una etapa presintomática. El diseño de estudio preferido es el estudio transversal (v. sección «Estudios transversales» y el cap. 8).
- *Pronóstico:* con el fin de determinar qué es probable que le suceda a alguien cuya enfermedad se detecta en una etapa temprana. El diseño de estudio preferido es el estudio longitudinal (v. sección «Estudios transversales»).
- *Causalidad:* se intenta determinar si un supuesto agente nocivo, como la contaminación del medio ambiente, se relaciona con el desarrollo de enfermedades. El diseño de estudio preferido es el estudio de cohortes o de casos y controles, dependiendo de lo infrecuente que sea la enfermedad (v. secciones «Estudios transversales» y «Publicaciones de casos aislados»), aunque las publicaciones de casos aislados (v. sección «Jerarquía tradicional de la evidencia») también pueden proporcionar información crucial.
- *Estudios psicométricos:* se realiza una medición de actitudes, creencias o preferencias, a menudo sobre la naturaleza de la enfermedad o su tratamiento.

Los estudios cualitativos se describen en el capítulo 12.

Para obtener más ayuda con esta tarea (que suele resultar difícil hasta que se domina), puede consultarse la página web del Oxford Centre for Evidence-Based Medicine (www.cebm.ox.ac.uk).

Ensayos controlados aleatorizados

En un ECA, los participantes en el ensayo se asignan de forma aleatoria (por un proceso equivalente a lanzar una moneda) a un grupo que recibe una intervención (p. ej., un tratamiento farmacológico) o a otro grupo (que recibe un placebo o, más frecuentemente, el mejor tratamiento disponible). Ambos grupos se siguen durante un período de tiempo predeterminado y se analizan en términos de resultados específicos definidos al inicio del estudio (p. ej., fallecimiento, infarto de miocardio y concentración sérica de colesterol). Puesto que, *por término medio,* los grupos son idénticos salvo para la intervención, todas las diferencias en los resultados son, en teoría, atribuibles a la intervención. En realidad, sin embargo, no todos los ECA son tan claros.

Algunos artículos que describen los ensayos donde se compara una intervención frente a un grupo de control no son realmente ensayos aleatorizados y se denominan *otros ensayos clínicos controlados,* un término empleado para describir los estudios comparativos en los que los participantes se asignaron a los grupos de intervención o control de una forma no aleatoria. Esta situación puede producirse, por ejemplo, cuando la asignación aleatoria fuese imposible, inviable o poco ética (p. ej., cuando los pacientes de la planta A reciben una dieta mientras que los de la planta B reciben otra dieta diferente). (Aunque este diseño es de peor calidad que el ECA, es mucho más fácil de llevar a cabo y se utilizó con éxito hace un siglo para demostrar el beneficio del arroz integral sobre el arroz blanco en el tratamiento del beriberi[8].) Los problemas de la asignación no aleatoria se describen con más detalle en la sección «¿Se ha evitado o minimizado el sesgo sistemático?» a la hora de determinar si los dos grupos de un ensayo se pueden comparar razonablemente entre sí a nivel estadístico.

Algunos ensayos son una especie de estudio intermedio entre los verdaderos ensayos aleatorizados y los no aleatorizados. En ellos, la aleatorización no se realiza verdaderamente al azar (p. ej., utilizando sobres sellados numerados secuencialmente, cada uno con un número aleatorio generado por ordenador en su interior), sino por algún método que permite al médico saber en qué grupo va a estar el paciente *antes de tomar una decisión definitiva de asignar de manera aleatoria al paciente.* Esto permite que se introduzcan sesgos sutiles, pues el médico podría ser más (o menos) tendente a incluir a un paciente particular en el ensayo si cree que dicho paciente recibiría el tratamiento activo. En especial, puede que los pacientes con una enfermedad más grave no se asignen inconscientemente al grupo de placebo del ensayo. Entre los ejemplos de métodos inaceptables se incluyen la asignación aleatoria por la última cifra de la fecha de nacimiento (pares al grupo A, impares al grupo B), el lanzamiento de una

Cuadro 3.3 Ventajas del diseño de ensayo controlado aleatorizado

1. Permite una evaluación rigurosa de una sola variable (p. ej., el efecto del tratamiento farmacológico frente al placebo) en un grupo de pacientes definido con precisión (p. ej., mujeres posmenopáusicas mayores de 50-60 años).
2. El diseño es prospectivo (es decir, se recogen datos de eventos que suceden *después* de que se decide realizar el estudio).
3. Utiliza el razonamiento hipotético-deductivo (es decir, pretende demostrar la falsedad, en lugar de confirmar su propia hipótesis; v. sección «Tres preguntas preliminares para orientarse»).
4. Erradica potencialmente el sesgo al comparar dos grupos por lo demás idénticos (pero v. el texto posterior y la sección «¿Se ha evitado o minimizado el sesgo sistemático?»).
5. Permite la realización de metaanálisis (combinación de los resultados numéricos de varios ensayos similares) con posterioridad; v. sección «Diez preguntas que deben plantearse sobre un artículo que pretende validar una prueba diagnóstica o de cribado»).

moneda (cara al grupo A, cruz al grupo B), la asignación secuencial (paciente A al grupo 1; paciente B al grupo 2, etc.) y la fecha en la que se ve al paciente en consulta (todos los pacientes vistos una semana al grupo 1 y todos los que se ven la semana siguiente al grupo 2, etc.) (cuadro 3.3)[9,10].

A continuación, se presentan varios ejemplos de preguntas clínicas que sería mejor responder mediante un ECA, pero deben tenerse en cuenta también los ejemplos de las secciones posteriores de este capítulo sobre las situaciones en las que podrían o deberían utilizarse otros tipos de estudios en su lugar:

- ¿Es este fármaco mejor que un placebo o que un fármaco diferente para una enfermedad en particular?
- ¿Es un nuevo procedimiento quirúrgico mejor que el método de elección actual?
- ¿Es un algoritmo de apoyo a la toma de decisiones *online* mejor que los consejos verbales para ayudar a los pacientes a tomar decisiones informadas sobre las opciones terapéuticas de una enfermedad particular?
- ¿Cambiar una dieta rica en grasas saturadas por una rica en grasas poliinsaturadas afectará significativamente a la concentración sérica de colesterol?

Los ECA a menudo se consideran el patrón oro de la investigación médica, lo que es cierto hasta cierto punto (v. sección «Jerarquía tradicional de la evidencia»), pero sólo para ciertos tipos de preguntas clínicas (v. cuadro 3.2 y las secciones «Estudios de cohortes», «Estudios de casos y controles», «Estudios transversales» y «Publicaciones de casos aislados»). Todas las preguntas que mejor se prestan al diseño de tipo ECA se refieren a *intervenciones* y tienen que ver principalmente

Cuadro 3.4 Inconvenientes del diseño de ensayo controlado aleatorizado

Caro y laborioso, por lo que, en la práctica:

- Muchos ECA o bien nunca se realizan o bien se llevan a cabo en muy pocos pacientes o durante un período demasiado corto (v. sección «¿Se abordaron las cuestiones estadísticas preliminares?»).
- La mayoría de los ECA están financiados por grandes organismos de investigación (patrocinados por universidades o por el gobierno) o por compañías farmacéuticas, que, en última instancia, determinan la agenda de investigación.
- Los criterios de valoración indirectos pueden no reflejar los resultados que son importantes para los pacientes (v. sección «Criterios de valoración indirectos»).

Puede introducir un «sesgo oculto», sobre todo debido a:

- Una asignación aleatoria imperfecta (v. texto anterior).
- La ausencia de asignación aleatoria de todos los pacientes elegibles (el médico sólo ofrece participar en el ensayo a los pacientes que considera que responderán bien a la intervención).
- El fallo a la hora de ocultar a los evaluadores el estatus de aleatorización de los pacientes (v. sección «¿Se realizó la evaluación de forma "ciega"?»).

con el tratamiento o la prevención. Sin embargo, hay que recordar que incluso a la hora de evaluar las intervenciones terapéuticas, y sobre todo en el caso contrario, los ensayos aleatorizados se asocian a varios inconvenientes significativos (v. cuadro 3.4)[11,12].

Hay que recodar también que los resultados de un ECA pueden tener una aplicabilidad limitada debido a los criterios de exclusión (reglas sobre quién puede quedar fuera del estudio), el sesgo de inclusión (selección de participantes en el ensayo a partir de un grupo que no es representativo de todas las personas que tienen la enfermedad [v. sección «¿Qué pacientes incluye el estudio?»]), el rechazo (o incapacidad) de ciertos grupos de pacientes a la hora de dar su consentimiento para ser incluidos en el estudio, el análisis exclusivamente de los criterios de valoración «objetivos» predefinidos que puede excluir aspectos cualitativos importantes de la intervención (v. cap. 12) y el sesgo de publicación (es decir, la publicación selectiva de resultados positivos, a menudo, pero no siempre, porque la organización que financió la investigación tiene mucho que ganar o perder en función de los resultados[9,10]). Por otra parte, los ECA pueden instrumentalizarse de forma correcta o incorrecta[2] y, una vez publicados, sus resultados pueden distorsionarse por una comunidad científica demasiado entusiasta o por un público ávido de un nuevo fármaco maravilloso[13]. Aunque todos estos problemas también pueden plantearse con otros diseños de ensayos, pueden ser particularmente

relevantes cuando un ECA se nos presenta como óptimo desde el punto de vista metodológico.

Además, hay muchas situaciones en las que los ECA son innecesarios, poco prácticos o inadecuados:

Los ECA son innecesarios:

- Cuando se descubre una intervención claramente satisfactoria para una enfermedad mortal si no se aplica.
- Cuando un ECA o metaanálisis previo ha ofrecido un resultado definitivo (tanto positivo como negativo; v. sección «Probabilidad y confianza»). Podría decirse que es *poco ético* asignar de forma aleatoria a los pacientes en un ensayo clínico sin antes realizar una revisión sistemática de la literatura para decidir si el ensayo debe llevarse a cabo.

Los ECA son poco prácticos:

- Cuando no sea ético solicitar el consentimiento para la asignación aleatoria (v. sección «Consideraciones éticas»).
- Cuando el número de participantes necesario para demostrar una diferencia significativa entre los grupos es demasiado alto (v. sección «¿Se abordaron las cuestiones estadísticas preliminares?»).

Los ECA son inapropiados:

- Cuando el estudio evalúa el pronóstico de una enfermedad. Para este análisis, el método adecuado para obtener la mejor evidencia es un estudio longitudinal de una *cohorte de inicio* correctamente organizada (v. sección «Estudios transversales»).
- Cuando el estudio evalúa la validez de una prueba diagnóstica o de cribado. Para este análisis, el método adecuado para obtener la mejor evidencia es un *estudio transversal* de pacientes con sospecha clínica de presentar el trastorno relevante (v. sección «Estudios transversales» y cap. 7).
- Cuando el estudio evalúa un asunto de «calidad asistencial» para el cual aún no se han establecido los criterios de «éxito». Por ejemplo, un ECA que compare métodos médicos frente a quirúrgicos de aborto podría evaluar el «éxito» en términos del número de pacientes en quienes se logró la evacuación completa, cuantía de la hemorragia y grado de dolor. Sin embargo, las pacientes podrían considerar que otros aspectos del procedimiento son importantes, como saber de antemano cuánto tiempo requerirá el procedimiento, no visualizar ni sentir el aborto, etcétera. Para este análisis, el método adecuado para obtener la mejor evidencia corresponde a los *métodos de investigación cualitativos* (v. cap. 12).

Todas estas cuestiones han sido debatidas exhaustivamente por los epidemiólogos clínicos, quienes nos recuerdan que menospreciar los ensayos no aleatorizados puede ser un signo de ingenuidad científica y no, como mucha gente suele asumir, de rigor intelectual[11]. También recomiendo al lector que preste atención a la ciencia emergente de los ECA *pragmáticos,* una metodología que tiene en cuenta las dificultades prácticas del mundo real para que los resultados de un ensayo sean más relevantes para dicho mundo real cuando se haya finalizado el ensayo[14]. Véase

también la sección «¿Qué información se puede esperar obtener de un artículo que describe un ensayo controlado aleatorizado: declaración CONSORT» donde describo las Consolidated Standards of Reporting Trials (CONSORT, Normas consolidadas para la publicación de ensayos clínicos) para presentar los resultados de los ECA.

Estudios de cohortes

En un estudio de cohortes, dos (o más) grupos de personas se seleccionan basándose en las diferencias en cuanto a su exposición a un agente específico (como una vacuna, un procedimiento quirúrgico o una toxina ambiental), tras lo cual se realiza un seguimiento para ver cuántas personas de cada grupo desarrollan una enfermedad, complicación u otro resultado en particular. El período de seguimiento en los estudios de cohortes suele medirse en años (y a veces en décadas) porque ése es el tiempo que muchas enfermedades, sobre todo el cáncer, tardan en aparecer. Hay que tener en cuenta que los ECA suelen iniciarse con personas que ya tienen una enfermedad, mientras que la mayoría de los estudios de cohortes se inician con personas que pueden desarrollar o no la enfermedad.

Un tipo especial de estudio de cohortes también se puede utilizar para determinar el pronóstico de una enfermedad (es decir, qué puede suceder a alguien que la padezca). Un grupo de personas en quienes se ha diagnosticado una enfermedad en una etapa precoz o con un resultado positivo en una prueba de cribado (v. cap. 7) se organiza (cohorte de inicio) y se sigue en repetidas ocasiones para determinar la incidencia (nuevos casos anuales) y la evolución temporal de diferentes resultados. (Estas definiciones deberían memorizarse si es posible: la *incidencia* es el número de nuevos casos anuales de una enfermedad, mientras que la *prevalencia* es la proporción global de la población que presenta la enfermedad.)

Sir Austen Bradford Hill, Sir Richard Doll y, recientemente, Sir Richard Peto llevaron a cabo el estudio de cohortes más famoso del mundo y sus autores fueron reconocidos con el título británico de caballero. Estos investigadores siguieron a 40.000 médicos varones británicos, divididos en cuatro cohortes (no fumadores, y fumadores leves, moderados y empedernidos) utilizando la mortalidad por todas las causas (cualquier fallecimiento) y por causas específicas (fallecimiento por una enfermedad particular) como medidas de resultado. La publicación de los resultados provisionales de un período de 10 años en 1964[15], que mostraron un exceso sustancial tanto de mortalidad por cáncer de pulmón como de mortalidad por cualquier causa en los fumadores, con una relación «dosis-respuesta» (es decir, cuanto más se fuma, mayor es la probabilidad de desarrollar cáncer de pulmón), contribuyó a demostrar que la relación entre el tabaquismo y la mala salud era causal en lugar de una coincidencia. Los resultados a los 20 años[16], 40 años[17] y 50 años[18] de este estudio trascendental (que logró un impresionante 94% de seguimiento de los participantes reclutados en 1951 y de quienes no había constancia de su fallecimiento) ilustran tanto los peligros de fumar como la solidez

de la evidencia que se puede obtener a partir de un estudio de cohortes realizado correctamente.

A continuación, se plantean preguntas clínicas que deberían abordarse mediante un estudio de cohortes:

- ¿Fumar causa cáncer de pulmón?
- ¿La píldora anticonceptiva «causa» cáncer de mama? (Debe tenerse en cuenta, una vez más, que la palabra «causa» es un término con muchas connotaciones y potencialmente engañoso. Como Guillebaud ha argumentado en su excelente libro «La píldora...»[19], si mil mujeres empezasen a tomar la píldora anticonceptiva oral mañana, algunas de ellas tendrían cáncer de mama, pero algunas de éstas lo habrían desarrollado de todas maneras. La pregunta que los epidemiólogos tratan de responder mediante los estudios de cohortes es: «¿cuál es el riesgo *adicional* de desarrollar cáncer de mama que correrían las mujeres por tomar la píldora, por encima del riesgo basal atribuible a su propio equilibrio hormonal, antecedentes familiares, dieta, consumo de alcohol, etc.?».)
- ¿La hipertensión arterial alta mejora con el tiempo?
- ¿Qué sucede con los lactantes que han nacido muy prematuramente, en términos de desarrollo físico y logros educativos posteriores?

Estudios de casos y controles

En un estudio de casos y controles, los pacientes con una enfermedad o trastorno particular se identifican y se «emparejan» con los controles (pacientes con alguna otra enfermedad, población general, vecinos o familiares). A continuación, se recogen datos (p. ej., mediante una búsqueda retrospectiva en las historias clínicas de estas personas o pidiéndoles que recuerden su propia historia) sobre la exposición previa a un posible agente causal de la enfermedad. Al igual que los estudios de cohortes, los de casos y controles suelen centrarse en la etiología de una enfermedad (es decir, su causa), en lugar de en su tratamiento. Se sitúan en un nivel más bajo en la jerarquía convencional de la evidencia (v. más adelante en el texto), pero este diseño suele ser la única opción para el estudio de las enfermedades raras. Una fuente importante de dificultades (y de posibles sesgos) en un estudio de casos y controles es la definición precisa de lo que se considera un «caso» porque la asignación incorrecta de una persona puede influir sustancialmente en los resultados (v. sección «¿Se evitó o se minimizó el sesgo sistemático?»). Además, este tipo de diseño no puede demostrar la causalidad, es decir, la *asociación* de A con B en un estudio de casos y controles no demuestra que A ha *causado* B.

A continuación, se enumeran algunas preguntas clínicas que deberían abordarse mediante un estudio de casos y controles:

- ¿El decúbito prono aumenta el riesgo de síndrome de muerte súbita del lactante?
- ¿La vacuna contra la tos ferina provoca lesión cerebral? (v. sección «¿Se evitó o se minimizó el sesgo sistemático?»).
- ¿Los tendidos de alta tensión causan leucemia?

Estudios transversales

Es probable que a todos se nos haya solicitado participar en un estudio, aunque sólo haya sido una encuestadora que nos haya preguntado en la calle cuál es la marca de pasta de dientes que preferimos. Los estudios que realizan los epidemiólogos se llevan a cabo básicamente siguiendo los mismos procedimientos: se recluta una muestra representativa de participantes y luego se entrevistan, se examinan o se estudian de otra forma para obtener respuestas a una pregunta clínica (o de otro tipo) específica. En los estudios transversales, los datos se recogen en un único punto temporal, pero pueden referirse retrospectivamente a experiencias de salud previas, por ejemplo, el estudio de las historias clínicas de los pacientes para ver con qué frecuencia se ha medido su presión arterial en los últimos 5 años.

Las siguientes preguntas clínicas deberían abordarse mediante un estudio transversal:

- ¿Cuál es la talla «normal» de un niño de 3 años? Esta pregunta, al igual que otras relativas al rango de la normalidad, se puede responder simplemente midiendo la altura de un número suficiente de niños sanos de 3 años. Sin embargo, este método no responde a la pregunta clínica relacionada: «¿cuándo se debe evaluar a un niño con una talla inusualmente baja para determinar si tiene una enfermedad?» porque, como en casi todas las mediciones biológicas, lo fisiológico (normal) se superpone con lo patológico (anormal). Este problema se plantea con más detalle en la sección «Cocientes de verosimilitudes».
- ¿Qué opinan las enfermeras psiquiátricas sobre la utilidad de los fármacos antidepresivos y de las terapias conversacionales para el tratamiento de la depresión grave?
- ¿Es cierto que «la mitad de todos los casos de diabetes están sin diagnosticar»? Éste es un ejemplo de la pregunta más general: «¿cuál es la prevalencia (proporción de personas con la enfermedad) de esta enfermedad en esta comunidad?». La única forma de obtener la respuesta es realizar la prueba diagnóstica definitiva en una muestra representativa de la población.

Publicaciones de casos aislados

En una publicación de un caso aislado, se describe la historia clínica de un paciente en forma de un relato («la paciente B es una secretaria de 54 años que presentó dolor torácico en junio de 2010...»). Las publicaciones de casos aislados suelen agruparse para constituir una *serie de casos* en la que se describen las historias clínicas de más de un paciente con una enfermedad particular para ilustrar un aspecto de ésta, el tratamiento o, más frecuentemente en la actualidad, una reacción adversa al tratamiento.

Aunque este tipo de investigación suele considerarse una evidencia científica relativamente débil (v. sección «Jerarquía tradicional de la evidencia»), la publicación de un caso aislado puede transmitir una gran cantidad de información que se perdería en un ensayo clínico o en un estudio (v. cap. 12). Además, los

médicos no académicos y el público profano comprenden con gran facilidad los casos publicados. Si es necesario, pueden escribirse y publicarse en pocos días, lo que les confiere una clara ventaja sobre los ensayos clínicos (cuyo período de gestación puede ser de varios años) o los metaanálisis (incluso más). Es indudable que existen buenas razones teóricas para volver a considerar la humilde publicación de casos aislados como una contribución útil y válida para la ciencia médica, entre otras razones porque un relato es una de las mejores maneras de *dar sentido* a una situación clínica compleja. Richard Smith, que fue editor del *British Medical Journal* durante 20 años, ha fundado recientemente una nueva revista llamada *Cases,* dedicada íntegramente a la publicación de casos clínicos aislados (v. http://casesjournal.com/casesjournal).

Las publicaciones de casos aislados serían un tipo de estudio apropiado en las siguientes situaciones clínicas:

- Un médico advierte que dos bebés nacidos en su hospital presentan ausencia de extremidades (focomelia). Ambas madres habían tomado un nuevo fármaco (talidomida) en la primera etapa del embarazo. El médico quiere alertar lo antes posible a sus colegas de todo el mundo de la posibilidad de daños relacionados con los fármacos[20]. (Cualquiera que se apresure a pensar que las publicaciones de casos aislados nunca están justificadas científicamente debería recordar este ejemplo.)
- Una paciente previamente sana desarrolla una peritonitis bacteriana espontánea (un problema infrecuente que un médico promedio sólo ve una vez cada 10 años). El equipo clínico responsable realiza una búsqueda en la literatura para obtener evidencia de investigación y elabora lo que considera un plan terapéutico basado en la evidencia. La paciente se recupera de forma satisfactoria. El equipo decide escribir este caso para que sirva de ejemplo a otros médicos (lo que se denomina publicación de un caso aislado basada en la evidencia[21]).

Jerarquía tradicional de la evidencia

La notación estándar del peso relativo que presentan los diferentes tipos de estudio primario a la hora de tomar decisiones sobre las intervenciones clínicas (la «jerarquía de la evidencia») los dispone en el siguiente orden:

1. Revisiones sistemáticas y metaanálisis (v. cap. 9).
2. ECA con resultados definitivos (es decir, los intervalos de confianza no se superponen con el umbral del efecto clínicamente significativo [v. sección «Probabilidad y confianza»]).
3. ECA con resultados no definitivos (es decir, una estimación puntual que sugiere un efecto clínicamente significativo, pero con intervalos de confianza que se superponen con el umbral para este efecto [v. sección «Probabilidad y confianza»]).
4. Estudios de cohortes.
5. Estudios de casos y controles.
6. Estudios transversales.

7. Publicaciones de casos aislados.

El culmen de la jerarquía se reserva, con toda justicia, para los artículos de investigación secundarios en los que se han recopilado todos los estudios primarios sobre un tema en particular y se ha efectuado su evaluación crítica según criterios rigurosos (v. cap. 9). Sin embargo, se debe tener en cuenta que ni siquiera el más férreo defensor de la MBE situaría un metaanálisis o un ECA poco cuidadoso que presentase defectos metodológicos graves por encima de un estudio de cohortes amplio y bien diseñado. Además, como se muestra en el capítulo 12, muchos estudios importantes y válidos en el campo de la investigación cualitativa no aparecen en absoluto en esta jerarquía particular de evidencia.

Dicho de otro modo, la evaluación de la contribución potencial de un estudio en particular a la ciencia médica requiere mucho más esfuerzo del que se necesita para ubicar su diseño en la escala previa de 7 puntos. Una publicación más reciente sobre las jerarquías de la evidencia sugiere que deberíamos evaluar los estudios según cuatro dimensiones: riesgo de sesgo, consistencia, aplicabilidad *(directness)* y precisión. Este enfoque complicaría cualquier pirámide sencilla de evidencia[22]. El corolario es que no se debe aplicar la jerarquía de la evidencia mecánicamente, sino que tan sólo es una regla general.

Varios de nosotros elaboramos una representación más compleja de la jerarquía de la evidencia orientada al ámbito de la pregunta (tratamiento/prevención, diagnóstico, perjuicio, pronóstico) en 2011[23], que puede descargarse en la página web del Centre for Evidence-Based Medicine (http://www.cebm.net/index.aspx?o = 5653). Sin embargo, antes de consultarla, el lector debería asegurarse de que no tiene dudas con la jerarquía tradicional (básica) descrita en esta sección.

Consideraciones éticas

En mi época de médico residente, conseguí un trabajo en un hospital universitario de renombre mundial. Una de mis humildes tareas era atender a los pacientes geriátricos (ancianos) en urgencias. Al poco tiempo, dos médicos elegantes y más experimentados que estaban buscando mi ayuda para su investigación (eso lo supe después) me invitaron a comer. A cambio de ver mi nombre entre los autores del artículo, debía realizar una biopsia rectal (es decir, cortar un pequeño fragmento de tejido del recto) en todos los pacientes mayores de 90 años con estreñimiento. Solicité una copia del formulario de consentimiento que debería entregarse a los pacientes para que firmasen. Cuando me aseguraron que el paciente promedio de 90 años apenas notaría el procedimiento, aquello no me gustó y me negué a colaborar con su proyecto.

En aquel momento, ignoraba ingenuamente la gravedad de la infracción planeada por estos médicos. Realizar *cualquier* investigación, sobre todo una que conlleve procedimientos invasivos, en pacientes vulnerables y enfermos sin tener en cuenta todas las cuestiones éticas es tanto un delito penal como un posible motivo para que un médico sea «inhabilitado» en su profesión. Obtener la aprobación ética formal para un estudio de investigación (los lectores de Reino Unido pueden

consultar corec.org.uk) y garantizar que la investigación se realiza y se supervisa adecuadamente (conjunto de tareas y responsabilidades denominado *control de la investigación*[24-26]) puede constituir un enorme obstáculo burocrático. Por desgracia, los aspectos éticos a veces se ignoraban en el pasado en las investigaciones con bebés, ancianos, personas con dificultades de aprendizaje y pacientes sin posibilidad de protestar (p. ej., presos y militares), lo que dio lugar a algunos escándalos tristemente conocidos en el campo de la investigación[24].

En la actualidad, la mayoría de los editores se niegan sistemáticamente a publicar una investigación que no haya sido aprobada por un comité ético de investigación. Sin embargo, debe tenerse en cuenta que los enfoques muy estrictos de control de la investigación por los organismos oficiales pueden ser éticamente cuestionables. El profesor Charles Warlow[27], neurólogo e investigador, afirmó hace unos años que el excesivo énfasis en el «consentimiento informado» por parte de los comités éticos de investigación bienintencionados ha cavado la tumba de la investigación sobre traumatismos craneoencefálicos, ictus y otros problemas cerebrales agudos (en los cuales, como es evidente, el paciente no está en condiciones de considerar los pros y los contras personales de participar en un estudio de investigación). Más recientemente, un grupo de investigadores exasperados publicó un relato catártico titulado *Bureaucracy stifles medical research in Britain* («La burocracia ahoga la investigación médica en Gran Bretaña»[28]). El mensaje principal que puede extraerse de este libro es: asegúrese de que el estudio que está leyendo ha obtenido la aprobación ética y a la vez solidarícese con los investigadores que han tenido que «hacer malabares» para conseguirla.

Bibliografía

1 Altman DG. The scandal of poor medical research. BMJ: British Medical Journal 1994;**308**(6924):283.
2 Altman DG. Poor-quality medical research. JAMA:The Journal of the American Medical Association 2002;**287**(21):2765-7.
3 Godlee F, Jefferson T, Callaham M, et al. *Peer review in health sciences*. London: BMJ Books; 2003.
4 Popper KR. *The logic of scientific discovery*. Abingdon, UK: Psychology Press; 2002.
5 Anon. Randomised trial of intravenous streptokinase, aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. (ISIS-2 Collaborative Group). Lancet 1988;**ii**:349-60.
6 Lee A, Joynt GM, Ho AM, et al. Tips for teachers of evidence-based medicine: making sense of decision analysis using a decision tree. Journal of General Internal Medicine 2009;**24**(5):642-8.
7 Drummond MF, Sculpher MJ, Torrance GW. *Methods for the economic evaluation of health care programs*. Oxford: Oxford University Press, 2005.
8 Fletcher W. Rice and beriberi: preliminary report of an experiment conducted at the Kuala Lumpur Lunatic Asylum. Lancet 1907;**1**:1776.
9 Sterne JA, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ: British Medical Journal 2001;**323**(7304):101.

10 Cuff A. Sources of Bias in Clinical Trials. 2013. http://applyingcriticality.wordpress.com/2013/06/19/sources-of-bias-in-clinical-trials/(accessed 26th June 2013).
11 Kaptchuk TJ. The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? Journal of Clinical Epidemiology 2001;**54**(6):541-9.
12 Berwick D. Broadening the view of evidence-based medicine. Quality and Safety in Health Care 2005;**14**(5):315-6.
13 McCormack J, Greenhalgh T. Seeing what you want to see in randomised controlled trials: versions and perversions of UKPDS data. United Kingdom prospective diabetes study. BMJ: British Medical Journal 2000;**320**(7251):1720-3.
14 Eldridge S. Pragmatic trials in primary health care: what, when and how? Family Practice 2010;**27**(6):591-2 doi: 10.1093/fampra/cmq099.
15 Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. BMJ: British Medical Journal 1964;**1**(5395):1399.
16 Doll R, Peto R. Mortality in relation to smoking: 20 years' observations on male British doctors. BMJ: British Medical Journal 1976;**2**(6051):1525.
17 Doll R, Peto R, Wheatley K, et al. Mortality in relation to smoking: 40 years' observations on male British doctors. BMJ: British Medical Journal 1994;**309**(6959):901-11.
18 Doll R, Peto R Boreham J, et al. Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ: British Medical Journal 2004;**328**(7455):1519.
19 Guillebaud J, MacGregor A. *The pill and other forms of hormonal contraception*. USA: Oxford University Press, 2009.
20 McBride WG. Thalidomide and congenital abnormalities. Lancet 1961;**2**:1358.
21 Soares-Weiser K, Paul M, Brezis M, et al. Evidence based case report. Antibiotic treatment for spontaneous bacterial peritonitis. BMJ: British Medical Journal 2002;**324**(7329):100-2.
22 Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5 grading the strength of a body of evidence when comparing medical interventions – agency for healthcare research and quality and the effective health-care program. Journal of Clinical Epidemiology 2010;**63**(5):513-23 doi: 10.1016/j.jclinepi.2009.03.009.
23 Howick J, Chalmers I, Glasziou P, et al. *The 2011 Oxford CEBM levels of evidence (introductory document)*. Oxford: Oxford Centre for Evidence-Based Medicine, 2011.
24 Slowther A, Boynton P, Shaw S. Research governance: ethical issues. JRSM: Journal of the Royal Society of Medicine 2006;**99**(2):65-72.
25 Shaw S, Boynton PM, Greenhalgh T. Research governance: where did it come from, what does it mean? JRSM: Journal of the Royal Society of Medicine 2005;**98**(11):496-502.
26 Shaw S, Barrett G. Research governance: regulating risk and reducing harm? JRSM: Journal of the Royal Society of Medicine 2006;**99**(1):14-9.
27 Warlow C. Over-regulation of clinical research: a threat to public health. Clinical Medicine 2005;**5**(1):33-8.
28 Snooks H, Hutchings H, Seagrove A, et al. Bureaucracy stifles medical research in Britain: a tale of three trials. BMC Medical Research Methodology 2012;**12**(1):122.

Capítulo 4 Evaluación de la calidad metodológica

Como ya expuse en la sección «La ciencia de criticar los artículos», un artículo se hundirá o flotará dependiendo de la solidez de su sección de métodos. En este capítulo se analizan cinco preguntas esenciales en las que debemos basarnos a la hora de decidir si el artículo «se tira a la basura» de inmediato (por defectos metodológicos inexcusables), si se interpretan sus resultados con cautela (porque los métodos eran subóptimos) o si se confía en él por completo (porque no se observa ningún defecto de método). Estas cinco preguntas (era el estudio original, qué pacientes incluye el estudio, estaba bien diseñado, se evitó el sesgo sistemático [es decir, el estudio fue «controlado» de manera adecuada] y era lo bastante extenso y se continuó durante el tiempo suficiente para que los resultados sean creíbles) se analizarán de forma individual.

¿Era el estudio original?

En teoría, no tiene sentido probar una hipótesis científica que alguien más ya haya probado de una u otra manera. Sin embargo, en la vida real, la ciencia pocas veces es tan clara y concreta. Sólo una pequeña proporción de la investigación médica surca caminos totalmente nuevos y una proporción igual de pequeña repite exactamente los pasos de investigadores previos. La mayoría de los estudios de investigación informarán (si son metodológicamente sólidos) de que una hipótesis particular tiene ligeramente más o menos probabilidades de ser correcta que antes de haber añadido nuestra pieza a un rompecabezas más amplio. Por lo tanto, puede ser perfectamente válido realizar un estudio que sea, aparentemente, «no original». De hecho, toda la ciencia del metaanálisis depende de que haya más de un estudio en la literatura que haya abordado la misma cuestión más o menos de la misma manera.

Por lo tanto, la cuestión práctica que se debe plantear acerca de un nuevo trabajo de investigación no es «¿alguien ha realizado un estudio similar antes?», sino «¿esta nueva investigación añade algo a la literatura?». A continuación, se presenta una lista de ejemplos de este tipo:

- ¿Éste es un estudio más extenso, seguido durante más tiempo o más sustancial en algún otro sentido que los previos?

- ¿Los métodos de este estudio son más rigurosos? (En particular, ¿tiene en cuenta alguna de las críticas metodológicas específicas de los estudios anteriores?)
- ¿Los resultados numéricos de este estudio suponen una adición significativa a un metaanálisis de estudios previos?
- ¿Es la población estudiada diferente en algún aspecto? (p. ej., ¿el estudio ha evaluado diferentes grupos étnicos, edades o sexos que los estudios previos?)
- ¿La cuestión clínica planteada tiene la suficiente importancia, y existen suficientes interrogantes en la mente de quienes toman las decisiones públicas o fundamentales, para que la nueva evidencia sea «políticamente» deseable aunque no sea estrictamente necesaria desde el punto de vista científico?

¿Qué pacientes incluye el estudio?

Uno de los primeros artículos que me llamó la atención se titulaba «But will it help my patients with myocardial infarction?» («¿Pero esto ayudará a mis pacientes con infarto de miocardio?»)[1]. No recuerdo los detalles del artículo, pero me abrió los ojos al hecho de que la investigación sobre los pacientes de otro médico tal vez no aporte información para mi práctica personal. Esto no es mera xenofobia. Las principales razones por las cuales los participantes (Sir Iain Chalmers ha insistido enérgicamente en que no se les llame «pacientes»)[2] de un ensayo clínico o estudio podrían diferir de los pacientes de la «vida real» se enumeran a continuación:

(a) Estaban más o menos enfermos que los pacientes a los que tratamos.
(b) Eran de un grupo étnico diferente o llevaban un estilo de vida distinto del de nuestros propios pacientes.
(c) Recibieron más atención (o diferente) durante el estudio de la que podríamos proporcionar a nuestros pacientes.
(d) A diferencia de la mayoría de los pacientes de la vida real, no tenían ninguna otra enfermedad aparte de la afección que se estaba estudiando.
(e) Ninguno de ellos fumaba, bebía alcohol o tomaba la píldora anticonceptiva.

Por lo tanto, antes de aceptar los resultados de cualquier artículo, hay que plantearse las siguientes preguntas:

1. *¿Cómo se reclutó a los participantes?* Si quisiéramos realizar una encuesta sobre las opiniones de los usuarios del servicio de urgencias del hospital, se podría reclutar a los encuestados poniendo un anuncio en el periódico local. Sin embargo, este método sería un buen ejemplo de *sesgo de reclutamiento* ya que la muestra que se obtendría estaría sesgada a favor de los usuarios que tuviesen una gran motivación para responder a las preguntas y de aquellos a quienes les gustase leer periódicos. Sería mejor entregar un cuestionario a todos los usuarios (o a una muestra de uno de cada diez usuarios) que acudiesen un día en especial.
2. *¿Quién fue incluido en el estudio?* En el pasado, los ensayos clínicos solían excluir a las personas con enfermedades concomitantes, a quienes no hablaban el idioma del país, a quienes tomaban otros fármacos y a quien no sabía leer el formulario de consentimiento. Este método puede ser experimentalmente

adecuado, pero como los resultados de los ensayos clínicos se utilizarán para orientar la práctica en relación con grupos de pacientes más amplios, en realidad presenta defectos de índole científica. Es evidente que los resultados de los estudios farmacocinéticos de nuevos fármacos en varones voluntarios sanos de 23 años no serán aplicables a una anciana promedio. Este problema, que ha atormentado a algunos médicos y científicos durante décadas, ha sido asumido más recientemente por los propios pacientes, sobre todo por parte de los grupos de apoyo a pacientes que han solicitado la ampliación de los criterios de inclusión en los ensayos de fármacos contra el sida[3].

3. *¿Quién fue excluido del estudio?* Por ejemplo, un ensayo controlado aleatorizado puede estar restringido a los pacientes con formas moderadas o graves de una enfermedad, como la insuficiencia cardíaca, lo que podría dar lugar a conclusiones falsas sobre el tratamiento de la insuficiencia cardíaca *leve*. Esto tiene implicaciones prácticas importantes cuando los ensayos clínicos realizados en pacientes de las consultas externas de un hospital se utilizan para determinar cuál es la «mejor práctica» en atención primaria, donde el espectro de la enfermedad suele ser más leve.
4. *¿Se estudió a los participantes en circunstancias de la «vida real»?* Por ejemplo, ¿fueron ingresados en el hospital sólo para observación? ¿Recibieron explicaciones extensas y detalladas de los beneficios potenciales de la intervención? ¿Se les dio el número de teléfono de un investigador principal? ¿La empresa que financió la investigación proporciona nuevos equipos que no estarían disponibles para el médico habitual? Estos factores no invalidarían el estudio, pero pueden arrojar dudas sobre la aplicabilidad de sus resultados a nuestra propia práctica.

¿El diseño del estudio era acertado?

Aunque la terminología del diseño del ensayo de investigación puede resultar poco comprensible, gran parte de lo que se denomina de forma grandilocuente *valoración crítica* es puro sentido común. Personalmente, evalúo el diseño básico de un ensayo clínico con estas dos preguntas:

¿Qué intervención específica u otra actuación se estaba evaluando, y con qué se estaba comparando? Ésta es una de las preguntas más fundamentales a la hora de valorar cualquier artículo. Resulta tentador asumir las afirmaciones publicadas por su valor aparente, pero debe recordarse que los autores a menudo tergiversan (por lo general inconscientemente y no de forma deliberada) lo que realmente hicieron y sobrestiman su originalidad e importancia potencial. En los ejemplos de la tabla 4.1 he utilizado afirmaciones hipotéticas para no ofender, pero todas ellas se basan en errores similares observados en artículos publicados.

¿Qué resultado se midió y cómo? Si tuviésemos una enfermedad incurable para la que una compañía farmacéutica afirmase haber desarrollado un nuevo fármaco maravilloso, mediríamos la eficacia del fármaco en términos de si

Tabla 4.1 Ejemplos de descripciones problemáticas en la sección de métodos de un artículo

Afirmación de los autores	Lo que deberían haber dicho (o hecho)	Es un ejemplo de
«Hemos medido con qué frecuencia los médicos de cabecera preguntan a los pacientes si fuman»	«Revisamos las historias clínicas de los pacientes y determinamos en cuántos se había anotado su estatus de fumador»	Suposición de que las historias clínicas tienen una precisión del 100%
«Hemos evaluado cómo los médicos tratan la lumbalgia»	«Evaluamos lo que los médicos *dicen* que hacen al atender a un paciente con lumbalgia»	Suposición de que lo que los médicos dicen refleja lo que hacen en realidad
«Hemos comparado un tratamiento sustitutivo con parche de nicotina frente a un placebo»	«Se pidió a los participantes del grupo de intervención que se aplicasen un parche de 15 mg de nicotina dos veces al día; los del grupo control recibieron parches de aspecto idéntico»	No se ofrece información sobre la dosis del fármaco o el tipo de placebo
«Hemos pedido a 100 adolescentes que participasen en nuestro estudio sobre conductas sexuales»	«Contactamos con 147 adolescentes estadounidenses de raza blanca de 12-18 años (85 varones) en un campamento de verano; 100 de ellos (31 varones) aceptaron participar»	No se ofrece suficiente información sobre los participantes. (Obsérvese que en este ejemplo las cifras indican un sesgo de reclutamiento hacia las mujeres)
«Distribuimos a los pacientes de forma aleatoria para recibir "un plan de atención individual" o "la atención habitual"»	«El grupo de intervención recibió un plan de atención individual consistente en……; los pacientes del grupo control recibieron……»	No se ofrece suficiente información sobre la intervención. (Debería ofrecerse bastante información para permitir que otros investigadores repitiesen el estudio)
«Para evaluar la utilidad de un folleto informativo, dimos al grupo de intervención un folleto y un número de atención telefónica. El grupo control no recibió ninguno de los dos»	Si el estudio simplemente trata de evaluar la utilidad del folleto, ambos grupos deberían haber recibido el número de atención telefónica	No se trataron los grupos del mismo modo salvo por la intervención específica
«Hemos evaluado el uso de la vitamina C en la prevención del resfriado común»	«Una búsqueda sistemática de la literatura habría encontrado muchos estudios previos sobre el tema (v. sección «¿Cuándo es sistemática una revisión?»)	Estudio no original

nos prolongaría la vida (y, tal vez, de si *valdría la pena* vivir debido a nuestra situación y a los efectos secundarios del fármaco). Estaríamos poco interesados por las concentraciones sanguíneas de ciertas enzimas desconocidas que el fabricante señala como un indicador fiable de nuestras posibilidades de supervivencia. El uso de estos *criterios de valoración indirectos* se comenta en la sección «Criterios de valoración indirectos» en la página 81.

La medición de los efectos sintomáticos (p. ej., dolor), funcionales (p. ej., movilidad), psicológicos (p. ej., ansiedad) o sociales (p. ej., incomodidad) de una intervención plantea aún más problemas. La metodología del desarrollo, administración e interpretación de estas medidas de resultado subjetivas está fuera del alcance de este libro. Sin embargo, en general, siempre se debe buscar en el artículo la evidencia de que la medida de resultado se ha validado objetivamente, es decir, que alguien ha demostrado que la «medida de resultado» utilizada en el estudio mide lo que pretende medir y que los cambios de esta medida de resultado reflejan adecuadamente las variaciones del estado del paciente. Debe recordarse que lo importante en opinión del médico tal vez no tenga el mismo valor para el paciente y viceversa. Uno de los avances más interesantes de la medicina basada en la evidencia (MBE) en los últimos años es la ciencia emergente de las medidas de resultados referidos por los pacientes, que se describen en la sección «PROM» en la página 223.

¿Se evitó o se minimizó el sesgo sistemático?

El *sesgo sistemático* se define por parte de los epidemiólogos como cualquier elemento que influye erróneamente en las conclusiones acerca de los grupos y distorsiona las comparaciones[4]. Si el diseño de un estudio es un ensayo controlado aleatorizado (ECA), un ensayo comparativo no aleatorizado, un estudio de cohortes o un estudio de casos y controles, el objetivo debería ser que los grupos que se están comparando fuesen lo más parecidos entre sí como fuera posible, salvo por la diferencia específica que se está evaluando. Estos grupos deberían, en la medida de lo posible, recibir las mismas explicaciones, tener los mismos contactos con profesionales sanitarios y ser evaluados el mismo número de veces por los mismos evaluadores, utilizando las mismas medidas de resultado[5,6]. Los diferentes diseños de estudio emplean distintos métodos para reducir el sesgo sistemático.

Ensayos clínicos controlados aleatorizados

En un ECA, el sesgo sistemático se evita (en teoría) mediante la selección de una muestra de participantes de una población en particular y su asignación aleatoria a los diferentes grupos. En la sección «Ensayos controlados aleatorizados» se describen algunas formas en las que el sesgo puede introducirse incluso en este patrón oro de diseño de ensayos clínicos y en la figura 4.1 se resumen las fuentes específicas que deben comprobarse.

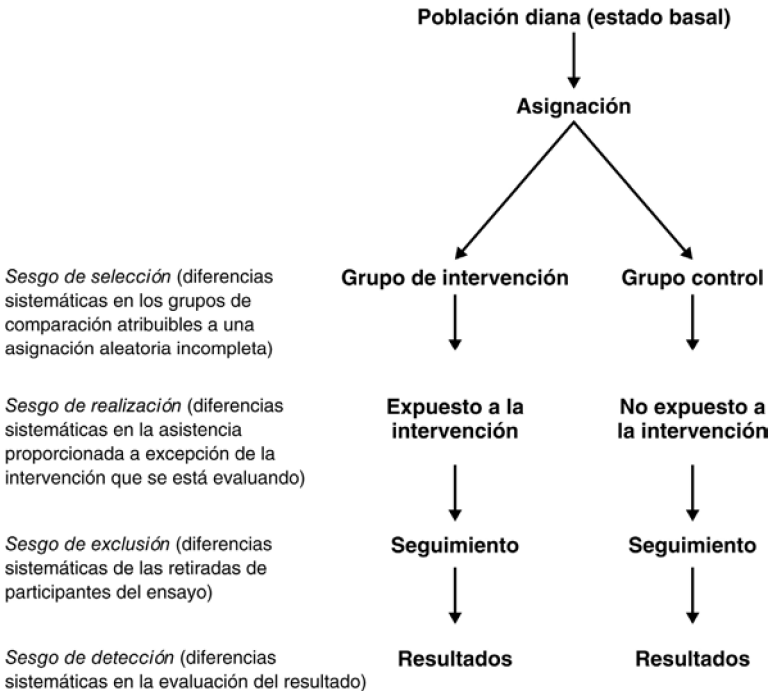

Figura 4.1 Fuentes de sesgo que deben comprobarse en un ensayo controlado aleatorizado.

Ensayos clínicos controlados no aleatorizados

Una vez presidí un seminario en el que un grupo multidisciplinario de estudiantes de medicina, enfermería, farmacia y profesiones asociadas presentaron los resultados de varios estudios de investigación realizados en nuestro centro. Todos los estudios presentados menos uno, tenían un diseño comparativo, pero no aleatorizado, es decir, un grupo de pacientes (p. ej., pacientes de consultas externas de hospital con asma) habían recibido una intervención (p. ej., un folleto informativo), mientras que otro grupo (p. ej., los pacientes que consultaban con su médico general por asma) había recibido otra intervención (p. ej., sesiones educativas de grupo). Me sorprendió el número de ponentes que creían que su estudio era un ECA o equivalente a él. Dicho de otro modo, estos jóvenes investigadores, encomiablemente entusiastas y comprometidos, no habían advertido el sesgo más evidente de todos: estaban comparando dos grupos que tenían diferencias inherentes y autoseleccionadas incluso antes de que se aplicase la intervención (además de contar con todas las fuentes potenciales adicionales de sesgo de los ECA que se recogen en la fig. 4.1).

Como regla general, si el artículo que estamos leyendo es un ensayo clínico controlado no aleatorizado, debemos usar el sentido común para decidir si es probable que las diferencias iniciales entre los grupos de intervención y control eran tan grandes como para invalidar cualquier diferencia atribuida a los efectos de la intervención. De hecho, esto es lo que sucede casi siempre[7]. A veces, los autores de un artículo de este tipo presentan las características relevantes de cada grupo (como el promedio de edad, la proporción de sexos y los marcadores de la gravedad de la enfermedad) en una tabla para que el lector pueda comparar estas diferencias por sí mismo.

Estudios de cohortes

La selección de un grupo de control comparable es una de las decisiones más difíciles a las que se enfrentan los autores de un estudio observacional (de cohortes o de casos y controles). Por ejemplo, pocos estudios de cohortes (a veces ninguno) logran identificar dos grupos de individuos que sean iguales en cuanto a edad, proporción de sexos, nivel socioeconómico, presencia de enfermedades coexistentes, etc., y que tengan como única diferencia su exposición al agente que se va a evaluar. En la práctica, gran parte del «control» en los estudios de cohortes se produce en la etapa de análisis, donde se realiza un ajuste estadístico complejo para las diferencias iniciales respecto a las variables clave. Si esto no se lleva a cabo adecuadamente, las pruebas estadísticas de probabilidad y los intervalos de confianza (v. sección «Probabilidad y confianza») serán peligrosamente engañosos[6,7].

Este problema se ilustra en los diversos estudios de cohortes realizados sobre los riesgos y beneficios del alcohol, que han demostrado constantemente una relación en forma de J entre el consumo de alcohol y la mortalidad. El mejor resultado (en términos de muerte prematura) corresponde al grupo de la cohorte con una ingesta moderada de alcohol[8]. Las personas que se declaran abstemias parecen tener una probabilidad significativamente mayor de fallecer jóvenes que la persona promedio que ingiere tres o cuatro bebidas alcohólicas al día.

Ahora bien, surge la duda de si se puede suponer que los abstemios son, *como promedio*, idénticos a los bebedores moderados salvo por la cantidad de alcohol ingerida. La respuesta es, evidentemente, que no. Como todos sabemos, la población de abstemios incluye a personas a quienes se ha prescrito dejar el alcohol por razones de salud («ex bebedores enfermos»), a individuos que por motivos de salud o de otro tipo han eliminado varios elementos adicionales de su dieta y estilo de vida, a miembros de ciertos grupos religiosos o étnicos que estarían infrarrepresentados en las otras cohortes (en particular, musulmanes y adventistas del séptimo día) y también a quienes beben como esponjas, pero optan por mentir al respecto.

Los detalles del modo en que los epidemiólogos controlaron estos distintos grados de ser abstemio se exponen en otras publicaciones[8,9]. Curiosamente, cuando estaba escribiendo la tercera edición de este libro en 2005, en aquel momento se aceptaba que, incluso cuando se realizaban ajustes en el análisis de las posibles variables de confusión en las personas que se describían a sí mismas como no

bebedores, el aumento del riesgo de mortalidad prematura en ellas persistía (es decir, la curva en J era un fenómeno auténtico)[8].

Sin embargo, cuando escribí la cuarta edición en 2010, se había publicado un análisis más sofisticado de varios estudios de cohortes (es decir, con un control más detallado de los «ex bebedores enfermos»)[9]. Este análisis demostró que, a igualdad de los demás factores, los abstemios no son más propensos a contraer cardiopatías que los bebedores moderados (de modo que la famosa «curva en J» puede haber sido un artefacto todo este tiempo). Posteriormente, un nuevo metaanálisis pretendía demostrar que la curva en J era un fenómeno auténtico y que el alcohol era en realidad protector en pequeñas cantidades[10], pero un año después, un nuevo análisis de los mismos estudios primarios llegó a la conclusión opuesta, al haber otorgado más peso a los denominados defectos metodológicos[11]. Esto podría ser un tema de conversación alrededor de una cerveza con los colegas interesados en la MBE.

Estudios de casos y controles

En los estudios de casos y controles (en los cuales, como ya he explicado en la sección «Publicaciones de casos aislados», se analizan de forma retrospectiva las experiencias de las personas con y sin una enfermedad en particular para identificar la exposición a posibles causas de dicha enfermedad), el proceso más expuesto al sesgo no es la evaluación del resultado sino el diagnóstico de lo que constituye un caso y la decisión de *cuándo* el individuo se convierte en un caso.

Un buen ejemplo de esto ocurrió hace unos años cuando se emprendieron acciones legales contra los fabricantes de la vacuna contra la tos ferina, que presuntamente había causado lesiones neurológicas en varios recién nacidos[12]. Con el fin de responder a la pregunta: «¿la vacuna provocó lesiones cerebrales?», se realizó un estudio de casos y controles en el que un «caso» se definió como un lactante previamente sano que había presentado crisis comiciales u otros signos sugestivos de lesión cerebral en la semana posterior a la vacunación. Un control era un lactante de la misma edad y sexo tomado del mismo registro de vacunación que había recibido la vacuna y que podía o no haber desarrollado síntomas en algún momento.

La nueva aparición de signos de lesión cerebral en lactantes aparentemente sanos es excepcional, pero sucede, y la relación con una vacunación reciente podría ser una coincidencia. Además, la mayor ansiedad de la opinión pública sobre el tema podría haber sesgado el recuerdo de los progenitores y de los profesionales sanitarios, de modo que aquellos lactantes cuyos síntomas neurológicos hubiesen precedido o aparecido algún tiempo después de la administración de la vacuna contra la tos ferina podrían haberse clasificado erróneamente como casos. El juez de este proceso dictaminó que la mala clasificación de tres de estos lactantes como «casos» en lugar de como controles dio lugar a la sobrestimación de las lesiones atribuibles a la vacuna contra la tos ferina, triplicando la incidencia real[12]. Aunque esta decisión se ha cuestionado con posterioridad, el principio sigue siendo válido (que la asignación de lo que constituye un caso en un estudio de casos y controles debe realizarse con rigor y objetividad si se quiere evitar el sesgo sistemático).

¿Se realizó la evaluación de forma «ciega»?

Incluso el intento más riguroso para lograr un grupo de control comparable será en vano si las personas que evalúan los resultados (p. ej., quienes juzgan si alguien aún presenta datos clínicos de insuficiencia cardíaca o quienes valoran si una radiografía ha «mejorado» desde la previa) saben a qué grupo se asignó el paciente al que están evaluando. Quien piense que la evaluación de los signos clínicos y la interpretación de las pruebas diagnósticas como el ECG y las radiografías es 100% objetiva, no sabe de qué va la historia[13].

En el capítulo «The Clinical Examination» («La exploración física») del libro *Clinical epidemiology: a basic science for clinical medicine* de Sackett y cols.[14] se proporciona una evidencia considerable de que los médicos encuentran lo que esperan y desean encontrar al explorar a los pacientes. Es infrecuente que dos médicos competentes lleguen a un acuerdo completo sobre cualquier aspecto determinado de la exploración física o la interpretación de cualquier prueba diagnóstica. El grado de acuerdo más allá del azar entre dos observadores se puede expresar matemáticamente como la puntuación kappa, donde una puntuación de 1,0 indica un acuerdo perfecto. Las puntuaciones kappa de diversos especialistas a la hora de evaluar la altura de la presión venosa yugular de un paciente, de clasificar la retinopatía diabética a partir de fotografías de la retina y de interpretar una mamografía fueron, respectivamente, 0,42, 0,55 y 0,67[14].

Esta digresión sobre la discordancia clínica debería haber convencido al lector de que los esfuerzos para mantener a los evaluadores «ciegos» (o, para no ofender a los discapacitados visuales, *enmascarados*) respecto a la asignación de sus pacientes en cada grupo no es en absoluto superflua. Si, por ejemplo, supiésemos que un paciente había sido asignado aleatoriamente para recibir un fármaco con actividad hipotensiva en lugar de un placebo, podríamos ser más tendentes a volver a comprobar una medición que fuese sorprendentemente alta. Éste es un ejemplo de *sesgo de realización,* que se presenta en la figura 4.1 junto con otras fuentes de error para el evaluador no ciego.

Un ejemplo excelente del control del sesgo mediante un «cegamiento» adecuado se publicó en la revista *Lancet* hace unos años[15]. Majeed y cols. realizaron un ECA que demostró, al contrario que los resultados de varios estudios anteriores, que el tiempo de recuperación (días en el hospital, días de baja laboral y tiempo necesario para reanudar la actividad completa) después de la extirpación laparoscópica de la vesícula biliar (abordaje de «cirugía mínimamente invasiva») no era más corto que el asociado con la operación abierta tradicional. La discrepancia entre este ensayo y sus predecesores puede haberse debido al intento meticuloso de los autores de reducir el sesgo (v. fig. 4.1). Los pacientes no fueron asignados al azar hasta después de la inducción de la anestesia general. Ni los pacientes ni sus cuidadores sabían qué operación se había realizado, ya que todos los pacientes abandonaban el quirófano con vendajes idénticos (incluso con manchas de sangre). Estos hallazgos obligaron a los autores anteriores a preguntarse si era el sesgo de expectativa (v. sección «Diez preguntas que deben plantearse sobre un

artículo que pretende validar una prueba diagnóstica o de cribado»), en lugar de la recuperación más rápida, lo que animó a los médicos a dar de alta antes a los pacientes del grupo de cirugía laparoscópica.

¿Se abordaron las cuestiones estadísticas preliminares?

Dado que yo no soy profesional de la estadística, sólo hay tres cifras que tiendo a comprobar en la sección de métodos de un artículo:

1. El tamaño muestral.
2. La duración del seguimiento.
3. El cumplimiento del seguimiento.

Tamaño muestral

Un prerrequisito esencial antes de embarcarse en un ensayo clínico es calcular el tamaño muestral («potencia»). Un ensayo debe ser lo bastante amplio para tener una alta probabilidad de detectar un efecto útil estadísticamente significativo si existe y, por lo tanto, que sea razonablemente seguro que no existe ningún beneficio si no se observa en el ensayo.

Para calcular el tamaño muestral, el médico debe decidir dos cosas:

- El nivel de diferencia entre los dos grupos que constituiría un *efecto clínicamente significativo*. Debe tenerse en cuenta que éste tal vez no coincida con el efecto estadísticamente significativo. Por citar un ejemplo de un ensayo clínico famoso sobre el tratamiento de la hipertensión, se podría administrar un nuevo fármaco que disminuyese la presión arterial alrededor de 10 mmHg y el efecto sería una reducción estadísticamente significativa de las posibilidades de sufrir un ictus (es decir, las posibilidades de que la menor incidencia se debiese al azar serían menores de 1/20)[16]. Sin embargo, si las personas a quienes se pide que tomen este fármaco sólo tuviesen una hipertensión arterial leve, sin otros factores de riesgo importantes de ictus (es decir, fueran relativamente jóvenes, no diabéticas, con cifras normales de colesterol, etc.), este nivel de diferencia sólo prevendría alrededor de un ictus en cada 850 pacientes tratados (una diferencia clínica respecto al riesgo que no compensaría a muchos pacientes por la molestia de tomar los comprimidos). Esto se demostró hace más de 20 años y se ha confirmado en numerosos estudios desde entonces (v. una revisión Cochrane reciente[17]). Sin embargo, demasiados médicos aún tratan a sus pacientes en función de la significación *estadística* de los resultados de megaensayos en lugar de por la significación clínica para sus pacientes; esto motiva (según afirman algunos) que ahora tengamos casi una epidemia de hipertensión leve sobretratada[18].
- La media y la desviación estándar (abreviada como DE; v. sección «¿Han planteado los autores correctamente el escenario?») de la variable de resultado principal.

Si el resultado en cuestión es un evento (como una histerectomía) en lugar de una cantidad (como la presión arterial), los datos necesarios son el porcentaje

de personas que presentaron el evento en la población y una estimación de lo que podría constituir un cambio clínicamente significativo de ese porcentaje.

Una vez que estos datos se han determinado, el tamaño muestral mínimo se puede calcular fácilmente usando fórmulas estándar, nomogramas o tablas, que se pueden obtener a partir de artículos publicados[19], manuales[20], sitios web de acceso gratuito (consulte http://www.macorr.com/ss_calculator.htm) o de paquetes de software estadístico comerciales (v., p. ej., http://www.ncss.com/pass.html). Por lo tanto, los investigadores pueden, *antes de que comience el ensayo*, averiguar el tamaño de la muestra que necesitan para tener una probabilidad moderada, alta o muy alta de detectar una verdadera diferencia entre los grupos. La probabilidad de detectar una diferencia verdadera se denomina *potencia* del estudio. Es habitual que los estudios estipulen una potencia del 80-90%. Por lo tanto, cuando se lee un artículo sobre un ECA, debería buscarse una frase que diga algo así (tomada del artículo de Majeed y cols. sobre la colecistectomía descrito antes)[15]:

> *Para tener el 90% de probabilidades de detectar una diferencia de una noche de ingreso en el hospital usando la prueba de U de Mann-Whitney (v. cap. 5, tabla 5.1), se necesitaban 100 pacientes en cada grupo (suponiendo una DE de 2 noches). Esto confiere una potencia mayor del 90% para detectar una diferencia de la duración de la intervención de 15 minutos, suponiendo una DE de 20 minutos.*

Si el artículo que se está leyendo no ofrece un cálculo del tamaño muestral *y* parece demostrar que no hay ninguna diferencia entre los grupos de intervención y de control del ensayo, se debería extraer del artículo (o directamente de los autores) la información de los puntos (a) y (b) previos y hacer el cálculo por nosotros mismos. La literatura médica está plagada de estudios de poca potencia, por lo general debido a que los autores tuvieron más dificultades de lo previsto para reclutar a sus participantes. Estos estudios suelen dar lugar a un error de tipo II o β, es decir, la conclusión errónea de que una intervención no tiene efecto. (En cambio, el tipo de error I, o α, que es más infrecuente, es la conclusión de que una diferencia es significativa cuando en realidad se debe a un error de muestreo.)

Duración del seguimiento

Aunque el propio tamaño muestral fuese adecuado, un estudio debe continuarse durante el tiempo suficiente para que el efecto de la intervención se refleje en la variable de resultado. Si los autores estaban evaluando el efecto de un nuevo analgésico sobre el grado de dolor postoperatorio, puede que su estudio sólo necesitase un período de seguimiento de 48 h. Por otro lado, si estaban analizando el efecto de la suplementación nutricional en la etapa preescolar sobre la talla adulta final, el seguimiento debería haberse medido en décadas.

Aunque la intervención haya demostrado una diferencia significativa entre los grupos después de 6 meses, por ejemplo, puede que esta diferencia no perdure. Como muchas personas que están a dieta saben por amarga experiencia, las

estrategias para reducir la obesidad suelen mostrar resultados espectaculares después de 2 o 3 semanas, pero si el seguimiento se prolonga durante un año o más, los participantes desafortunados (casi siempre) acaban por recuperar la mayor parte del peso.

Cumplimiento del seguimiento

Se ha demostrado en repetidas ocasiones que los participantes que abandonan los estudios de investigación son menos propensos a haber tomado sus comprimidos según las instrucciones, más propensos a no haber acudido a sus chequeos intermedios y más propensos a haber experimentado efectos secundarios con cualquier fármaco que quienes no lo abandonan. Las personas que no completan los cuestionarios pueden tener una opinión distinta sobre el tema (y probablemente menos interés) que quienes los devuelven cumplimentados por correo. Quienes participan en un programa de reducción de peso son más propensos a seguir volviendo si en realidad están perdiendo peso.

A continuación, se presentan algunas de las razones por las cuales los pacientes abandonan los ensayos clínicos (o son retirados de ellos por los investigadores):

1. Inclusión incorrecta del paciente en el ensayo (es decir, el investigador descubre durante el ensayo que el paciente ni siquiera debería haber sido asignado al azar porque no cumplía los criterios de inclusión).
2. Sospecha de reacciones adversas al fármaco del ensayo. Debe tenerse en cuenta que nunca hay que fijarse en la tasa de «reacciones adversas» del grupo de intervención sin compararla con la del placebo. Los comprimidos inertes provocan la aparición de una erupción con una frecuencia sorprendente.
3. Pérdida de motivación del participante («No quiero tomar estos comprimidos nunca más»).
4. Razones clínicas (p. ej., enfermedad concurrente, embarazo).
5. Pérdida del seguimiento (p. ej., el participante se muda de ciudad).
6. Fallecimiento. Evidentemente, las personas que mueren no asistirán a sus citas ambulatorias, por lo que, si no se tienen en cuenta específicamente, podrían clasificarse erróneamente como abandonos. Ésta es una razón por la cual los estudios con una tasa de seguimiento bajo (p. ej., menos del 70%) suelen considerarse poco fiables.

Si se ignora a todos los participantes que no han completado un ensayo clínico, se producirá un sesgo en los resultados, por lo general a favor de la intervención. Por lo tanto, la práctica estándar consiste en analizar los resultados de los estudios comparativos por *intención de tratar*. Esto significa que todos los datos de los participantes inicialmente asignados al grupo de intervención del estudio, incluidos los que abandonaron antes de su finalización, los que no tomaron el fármaco e incluso quienes posteriormente recibieron la intervención de control por cualquier motivo, deben analizarse junto con los datos de los pacientes que siguieron el protocolo en su totalidad. En cambio, los abandonos del grupo placebo del estudio deben analizarse con los que tomaron fielmente el placebo. Si leemos atentamente un artículo, se suele encontrar la frase, «los resultados se analizaron

por intención de tratar», pero no deberíamos quedarnos tranquilos hasta haber revisado y confirmado las cifras por nosotros mismos.

De hecho, hay algunas situaciones en las que el análisis por intención de tratar no se usa, con razón. La más frecuente es el *análisis de eficacia (o por protocolo),* que consiste en explicar los efectos de la propia intervención y que es, por lo tanto, el análisis del tratamiento recibido realmente. Sin embargo, incluso si los participantes en un análisis de eficacia forman parte de un ECA, para los fines del análisis constituyen realmente un estudio de cohortes (v. sección «Estudios de cohortes»).

Resumen

Después de haber analizado a fondo la sección de Métodos de un artículo, deberíamos ser capaces de plasmar por escrito brevemente qué tipo de estudio se ha realizado, con cuántos participantes, de dónde procedían éstos, qué tratamiento u otra intervención se aplicó, la duración del seguimiento (o, si se realizó una encuesta, cuál fue la tasa de respuesta) y qué medidas de resultado se utilizaron. En esta etapa, también deberíamos ser capaces de identificar qué pruebas estadísticas, en su caso, se emplearon para analizar los datos (v. cap. 5). Si tenemos claras estas cosas antes de leer el resto del artículo, los resultados nos resultarán más fáciles de entender, de interpretar y, llegado el caso, de rechazar. Deberíamos ser capaces de efectuar descripciones como las que se recogen a continuación:

Este artículo describe un ensayo aleatorizado no ciego sobre un tratamiento, realizado con 267 pacientes de consultas externas de hospital con un rango de edades de 58-93 años, en quienes se comparó el uso de un vendaje compresivo de cuatro capas frente a apósitos monocapa estándar para el tratamiento de las úlceras venosas no complicadas de las piernas. El seguimiento fue de 6 meses. El porcentaje de curación de la úlcera se midió desde la inclusión en el estudio en función de la superficie de un trazado de la herida realizado por una enfermera y calculado mediante un dispositivo de escaneo por ordenador. Los resultados se analizaron mediante el test de Wilcoxon de datos pareados.

Ésta es una encuesta realizada a 963 médicos generales seleccionados al azar en todo Reino Unido, a quienes se preguntó cuál fue su año de graduación en la facultad de medicina y la cifra a la que comenzarían el tratamiento de la hipertensión esencial. Las opciones de respuesta en el cuestionario estructurado fueron «por debajo de 89 mmHg», «90-99 mmHg» y «a partir de 100 mmHg».

Los resultados fueron analizados utilizando una prueba de chi cuadrado en una tabla de 3 × 2 para ver si el umbral para el tratamiento de la hipertensión se relacionaba con una fecha de graduación del médico de la facultad de medicina anterior o posterior a 1985.

Se trata de la publicación de un caso clínico aislado de un paciente con una sospecha de reacción adversa mortal a un fármaco, debida a un hipnótico recién comercializado.

Después de practicar la lectura de la sección de Métodos de los artículos de investigación siguiendo las indicaciones sugeridas en este capítulo, el lector encontrará que le faltará muy poco para empezar a utilizar las listas de comprobación del apéndice 1 o las indicaciones más exhaustivas de las Users' Guides to the Medical Literature (http://www.cche.net/usersguides/main.asp). Se volverá a insistir en muchos de los temas comentados aquí en el capítulo 6, en relación con la evaluación de los artículos que describen ensayos sobre el tratamiento farmacológico y otras intervenciones sencillas.

Bibliografía

1 Mitchell J. "But will it help my patients with myocardial infarction?". The implications of recent trials for everyday country folk British Medical Journal (Clinical Research Edition) 1982;**285**(6349):1140.
2 McCormack J, Greenhalgh T. Seeing what you want to see in randomised controlled trials: versions and perversions of UKPDS data. United Kingdom prospective diabetes study. BMJ: British Medical Journal 2000;**320**(7251):1720-3.
3 Phillips AN, Smith GD, Johnson MA. Will we ever know when to treat HIV infection? BMJ: British Medical Journal 1996;**313**(7057):608.
4 Coggon D, Barker D, Rose G. *Epidemiology for the uninitiated*. London: BMJ Books, 2009.
5 Cuff A. Sources of Bias in Clinical Trials. 2013. http://applyingcriticality.wordpress.com/2013/06/19/sources-of-bias-in-clinical-trials/(accessed 26th June 2013).
6 Delgado-Rodríguez M, Llorca J. Bias. Journal of Epidemiology and Community Health 2004;**58**(8):635-41 doi: 10.1136/jech.2003.008466.
7 Britton A, McKee M, Black N, et al. Choosing between randomised and non-randomised studies: a systematic review. Health Technology Assessment (Winchester, England) 1998;**2**(13): i.
8 Rimm EB, Williams P, Fosher K, et al. Moderate alcohol intake and lower risk of coronary heart disease: meta-analysis of effects on lipids and haemostatic factors. BMJ: British Medical Journal 1999;**319**(7224):1523.
9 Fillmore KM, Stockwell T, Chikritzhs T, et al. Moderate alcohol use and reduced mortality risk: systematic error in prospective studies and new hypotheses. Annals of Epidemiology 2007;**17**(5):S16-23.
10 Ronksley PE, Brien SE, Turner BJ, et al. Association of alcohol consumption with selected cardiovascular disease outcomes: a systematic review and meta-analysis. BMJ: British Medical Journal 2011;**342**:d671.
11 Stockwell T, Greer A, Fillmore K, et al. How good is the science? BMJ: BritishMedical Journal 2012;**344**:e2276.
12 Bowie C. Lessons from the pertussis vaccine court trial. Lancet 1990;**335**(8686):397-9.
13 Gawande A. *Complications: a surgeon's notes on an imperfect science*. London: Profile Books; 2010.

14 Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston, USA: Little, Brown and Company, 1985.
15 Majeed A, Troy G, Smythe A, et al. Randomised, prospective, single-blind comparison of laparoscopic versus small-incision cholecystectomy. The Lancet 1996;**347**(9007):989-94.
16 MRC Working Party. Medical Research Council trial of treatment of hypertension in older adults: principal results. BMJ: British Medical Journal 1992;**304**:405-12.
17 Diao D, Wright JM, Cundiff DK, et al. Pharmacotherapy for mild hypertension. Cochrane Database of Systematic Reviews (Online) 2012;**8** CD006742 doi: 10.1002/14651858.CD006742.pub2.
18 Spence D. Why do we overtreat hypertension? BMJ: British Medical Journal 2012;**345**:e5923 doi: 10.1136/bmj.e5923.
19 Charles P, Giraudeau B, Dechartres A, et al. Reporting of sample size calculation in randomised controlled trials: review. BMJ: British Medical Journal 2009;**338**:b1732.
20 Machin D, Campbell MJ, Tan S-B, et al. *Sample size tables for clinical studies*. Oxford: Wiley-Blackwell, 2011.

Capítulo 5 Estadística para no estadísticos

¿Cómo se pueden evaluar las pruebas estadísticas sin ser estadístico?

En esta era en la que la medicina se apoya cada vez más en las matemáticas, ningún clínico puede permitirse el lujo de dejar los aspectos estadísticos de un documento exclusivamente a los «expertos». A quien se considere, al igual que yo, un analfabeto en matemáticas, le recuerdo que no hay por qué ser capaz de construir un coche para poder conducir uno. Lo que se debe saber sobre las pruebas estadísticas es cuál de ellas es la mejor que podemos utilizar para los tipos habituales de cuestiones estadísticas. Hay que ser capaz de describir *con palabras* lo que hace la prueba y en qué circunstancias carece de validez o es inapropiada. En el cuadro 5.1 se muestran algunos trucos utilizados frecuentemente ante los que hay que estar alerta (en nuestra propia práctica y en la de otras personas).

La lista de comprobación resumida del apéndice 1, que se explica en detalle en las secciones siguientes, constituye mi propio método para evaluar la idoneidad de un análisis estadístico. Los lectores que la encuentren demasiado simplista pueden omitir esta sección y consultar la descripción más detallada para no estadísticos que aparece en la publicación *Basics Statistics for Clinicians* del *Canadian Medical Association Journal*[1-4] o los libros de estadística más convencionales. Los manuales de estadística preferidos por mis seguidores en Twitter se indican en la bibliografía[5-7]. Los lectores a quienes la estadística les resulte imposiblemente difícil deberían leer de uno en uno los puntos que se describen a continuación y pasar al siguiente sólo cuando crean haber comprendido por completo los anteriores. Ninguno de los puntos presupone un conocimiento detallado de los cálculos reales implicados.

La primera pregunta que debemos plantearnos es: «¿han utilizado los autores alguna prueba estadística?». Si en el artículo se presentan números y se afirma que dichos números tienen algún significado, pero no se han empleado métodos estadísticos para demostrarlo, es casi seguro que los autores han caminado sobre un terreno muy resbaladizo.

Cuadro 5.1 Diez formas de hacer trampa en las pruebas estadísticas al redactar los resultados

1. Introducir todos los datos en un ordenador y publicar como significativa cualquier relación con «$p < 0{,}05$» (v. sección «¿Se han calculado e interpretado adecuadamente los valores *p*?»).
2. Si las diferencias iniciales entre los grupos son favorables al grupo de intervención, recordar no realizar un ajuste en función de ellas (v. sección «¿Han determinado si sus grupos son comparables y, si es necesario, han realizado los ajustes en función de las diferencias iniciales?»).
3. No analizar los datos para comprobar si presentan una distribución normal. Si se hiciera, es posible que se deban utilizar pruebas no paramétricas, que no son tan sencillas (v. sección «¿Qué tipos de datos han recogido? ¿Se han utilizado las pruebas estadísticas apropiadas?»).
4. Ignorar todos los abandonos (exclusiones del estudio) y los no respondedores para que el análisis se refiera únicamente a los individuos que cumplieron plenamente con el tratamiento (v. sección «¿Se abordaron las cuestiones estadísticas preliminares?»).
5. Asumir siempre que se puede representar gráficamente un conjunto de datos frente a otro y calcular un «valor *r*» (coeficiente de correlación de Pearson) (v. sección «¿Se ha distinguido la correlación de la regresión, y se ha calculado e interpretado adecuadamente el coeficiente de correlación [valor *r*]?»), y que un valor *r* «significativo» demuestra la causalidad (v. sección «¿Se han realizado suposiciones sobre la naturaleza y la dirección de la causalidad?»).
6. Si los valores atípicos (puntos que están muy lejos de los otros en la gráfica) estropean los cálculos del estudio, limitarse a eliminarlos, pero si dichos valores atípicos son favorables para el estudio, aunque parezcan ser falsos resultados, dejarlos (v. sección «¿Se analizaron los valores atípicos con sentido común y se realizaron los ajustes estadísticos apropiados?»).
7. Si los intervalos de confianza del resultado se superponen con una diferencia nula entre los grupos, no incluirlos en la publicación. Mejor aún, mencionarlos brevemente en el texto, pero no representarlos en la gráfica e ignorarlos al extraer las conclusiones (v. sección «¿Se han calculado los intervalos de confianza, y se reflejan en las conclusiones de los autores?»).
8. Si la diferencia entre dos grupos se vuelve significativa a los 4,5 meses de un ensayo de 6 meses, interrumpir el ensayo y comenzar a escribir el artículo. Alternativamente, si a los 6 meses los resultados son «casi significativos», ampliar el ensayo otras 3 semanas (v. sección «¿Se han analizado los datos de acuerdo con el protocolo original del estudio?»).
9. Si los resultados resultan poco interesantes, repetir el análisis para comprobar si alguno de los subgrupos específicos se comportó de manera diferente. Quizá la intervención fuese eficaz en mujeres chinas de

52-61 años (v. sección «¿Se han analizado los datos de acuerdo con el protocolo original del estudio?»).

10. Si al analizar los datos de la forma prevista no se obtiene el resultado deseado, emplear otras pruebas estadísticas para analizar las cifras (v. sección «Si las pruebas estadísticas usadas en el artículo son poco claras, ¿por qué las escogieron los autores, e incluyeron una referencia?»).

¿Han planteado los autores correctamente el escenario?

¿Han determinado si sus grupos son comparables y, si es necesario, han realizado los ajustes en función de las diferencias iniciales?

La mayoría de los ensayos clínicos comparativos incluyen una tabla o un párrafo en el texto que muestra las características iniciales de los grupos estudiados (es decir, sus características *antes* de comenzar el ensayo o estudio observacional). Dicha tabla debería demostrar que los grupos de intervención y de control son similares en lo que respecta a la distribución por edades y sexo, así como a las variables pronósticas fundamentales (como el tamaño promedio de un tumor canceroso). Si hay diferencias importantes en estas características iniciales, aunque puedan deberse al azar, podrían dificultar la interpretación de los resultados. En tal caso, se pueden realizar ciertos ajustes para tratar de tenerlas en cuenta y, por lo tanto, reforzar las conclusiones del estudio. Para saber cómo se deben realizar los ajustes, puede consultarse la sección correspondiente en cualquiera de los principales libros de bioestadística, aunque no es necesario memorizar las fórmulas.

¿Qué tipo de datos se han recogido? ¿Se han utilizado las pruebas estadísticas apropiadas?

Los números suelen utilizarse para etiquetar las propiedades de las cosas. Se puede asignar un número para representar la talla, el peso, etcétera. Para las propiedades de este tipo, las mediciones pueden tratarse como números reales. Es posible, por ejemplo, calcular las medias del peso y la talla de un grupo de personas haciendo un promedio de las mediciones. En cambio, se puede poner un ejemplo diferente en el que se utilizan los números para etiquetar la propiedad «ciudad de origen», donde 1 corresponde a Londres, 2 a Manchester, 3 a Birmingham y así sucesivamente. En este caso, también podría calcularse la media de estos números para una muestra particular de casos, pero el resultado carecería de sentido. Lo mismo ocurriría si etiquetamos la propiedad «nivel de satisfacción por x», donde 1 corresponde a nada en absoluto, 2 a un poco y 3 a mucho. Aquí también podría calcularse la media del nivel de satisfacción, pero el resultado numérico no sería interpretable a menos que se supiese que la diferencia entre «nada en absoluto» y «un poco» fuese exactamente la misma que la diferencia entre «un poco» y «mucho».

Las pruebas estadísticas utilizadas en los artículos médicos suelen clasificarse como paramétricas (es decir, asumen que los datos se tomaron de un tipo especial de distribución, como una distribución normal) o no paramétricas (es decir, no asumen que los datos fueron tomados de un tipo particular de distribución).

Las pruebas no paramétricas se centran en el *orden de rango* de los valores (es decir, cuál es el más pequeño, cuál viene a continuación, etc.) e ignoran las diferencias absolutas entre ellos. Como se puede imaginar, es más difícil demostrar la significación estadística con las pruebas de orden de rango (de hecho, algunos estadísticos son escépticos sobre su utilidad), lo que ha llevado a que los investigadores utilicen medidas estadísticas como el valor *r* (v. sección «¿Se ha distinguido la correlación de la regresión, y se ha calculado e interpretado adecuadamente el coeficiente de correlación [valor *r*]?») de forma inapropiada. El valor *r* (paramétrico) no sólo es más fácil de calcular que una medida estadística de orden de rango equivalente como ρ (rho) de Spearman, sino que también es mucho más probable que proporcione (aparentemente) resultados significativos. Por desgracia, también proporcionará una estimación totalmente falsa y engañosa de la significación del resultado a menos que los datos sean apropiados para la prueba que se está utilizando. En la tabla 5.1 se presentan más ejemplos de pruebas paramétricas y sus equivalentes de orden de rango (si existen).

Otro aspecto que debe tenerse en cuenta es la forma de la distribución de la que se tomaron los datos. En mi época escolar, representábamos en mi clase la cuantía de la «paga» recibida en función del número de niños que recibían esa cantidad. Los resultados formaban un histograma similar al que se observa en la figura 5.1 y que constituye una distribución «normal». (El término *normal* se refiere a la forma de la gráfica y se utiliza debido a que muchos fenómenos biológicos muestran este patrón de distribución.) Algunas variables biológicas, como el peso corporal, tienen una distribución *asimétrica,* como se muestra en la figura 5.2. (En la figura 5.2 se muestra una asimetría negativa, mientras que el peso corporal presenta una asimetría positiva. La media del peso corporal de los varones adultos es de alrededor de 80 kg y existen personas que pesan 160 kg, pero nadie pesa menos de 0 kg, por lo que el gráfico no puede ser simétrico.)

Los datos que no siguen una distribución normal (asimétricos) a veces pueden transformarse para obtener un gráfico de forma normal representando el logaritmo de la variable asimétrica o realizando alguna otra transformación matemática (como la raíz cuadrada o el valor recíproco). Sin embargo, algunos datos no pueden transformarse en un patrón simétrico y la significación de esto se comenta más adelante. La decisión sobre si los datos presentan una distribución normal no es meramente académica, pues determinará el tipo de pruebas estadísticas que deben utilizarse. Por ejemplo, la regresión lineal (v. sección «Correlación, regresión y causalidad») ofrecerá unos resultados engañosos a menos que los puntos de la gráfica de dispersión formen una distribución particular alrededor de la línea de regresión, es decir, que los valores residuales (la distancia perpendicular desde cada punto a la línea) deben seguir a su vez una distribución normal. Transformar los datos para lograr una distribución normal (si fuese posible) no es hacer trampa.

Tabla 5.1 Pruebas estadísticas utilizadas con frecuencia

Prueba paramétrica	Ejemplo de prueba equivalente no paramétrica (orden de rango)	Objetivo de la prueba	Ejemplo
Prueba de *t* para dos muestras (no pareadas)	Prueba de *U* de Mann-Whitney	Compara dos muestras independientes extraídas de la misma población	Comparar las tallas de las niñas con las de los niños
Prueba de *t* para una muestra (pareada)	Prueba de datos pareados de Wilcoxon	Compara dos series de observaciones sobre una misma muestra (prueba la hipótesis de que la diferencia media entre dos mediciones es cero)	Comparar el peso de los lactantes antes y después de una toma
Análisis de la varianza unidireccional usando la suma total de los cuadrados (p. ej., prueba de *F*)	Análisis de la varianza por rangos (p. ej., prueba de Kruskall-Wallis)	Es una generalización de la prueba de *t* pareada o de la prueba de datos pareados de Wilcoxon donde se realizan tres o más series de observaciones en una única muestra	Determinar si la glucemia es mayor 1, 2 o 3 h después de una comida
Análisis de la varianza bidireccional	Análisis de la varianza bidireccional por rangos	Como se ha mencionado, pero prueba la influencia (e interacción) de dos covariables diferentes	En el ejemplo anterior, determinar si los resultados difieren entre varones y mujeres
No hay un equivalente directo	Prueba de χ^2	Prueba la hipótesis nula de que las proporciones de las variables estimadas a partir de dos (o más) muestras independientes son iguales	Evaluar si la aceptación en la facultad de medicina es más probable si el solicitante ha nacido en el país de la facultad
No hay un equivalente directo	Prueba de McNemar	Prueba la hipótesis nula de que las proporciones estimadas a partir de una muestra pareada son iguales	Comparar la sensibilidad y especificidad de dos pruebas diagnósticas diferentes cuando se aplican a la misma muestra

Coeficiente de correlación producto-momento (*r* de Pearson)	Coeficiente de correlación de rangos de Spearman (ρ)	Evalúa la *fuerza* de la asociación lineal entre dos variables continuas.	Evaluar si (y en qué medida) la concentración plasmática de HbA1 se relaciona con la concentración plasmática de triglicéridos en pacientes diabéticos
Regresión por el método de los mínimos cuadrados	No hay un equivalente directo	Describe la relación numérica entre dos variables cuantitativas, permitiendo predecir un valor a partir del otro	Observar en qué grado varía el flujo espiratorio máximo con la talla
Regresión múltiple por el método de los mínimos cuadrados	No hay un equivalente directo	Describe la relación numérica entre una variable dependiente y varias variables predictivas (covariables)	Determinar si (y en qué medida) la edad de una persona, la grasa corporal y la ingesta de sodio determinan su presión arterial

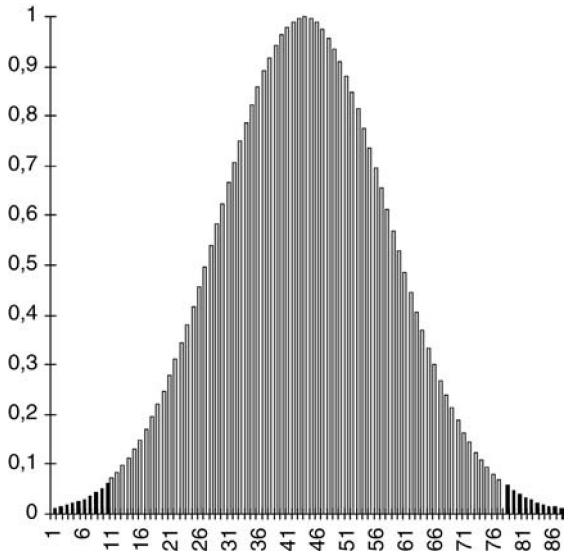

Figura 5.1 Ejemplo de una curva normal.

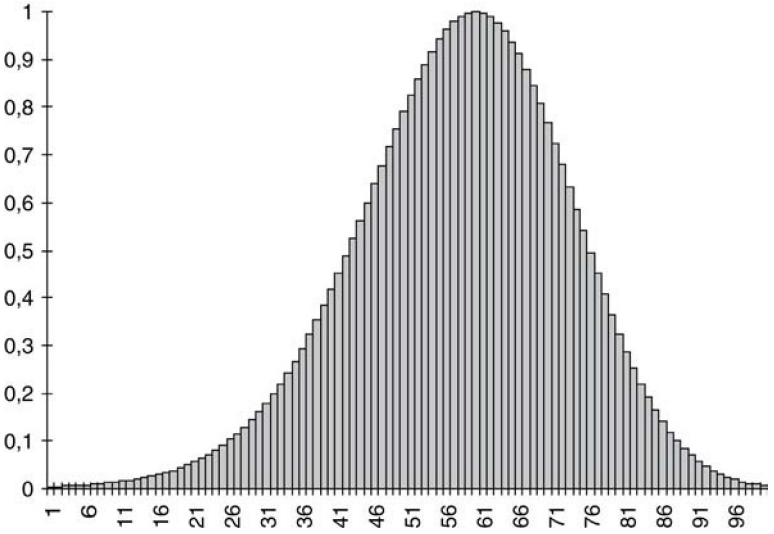

Figura 5.2 Ejemplo de una curva asimétrica.

Simplemente es la forma de asegurar que se otorga la importancia debida a los valores de datos a la hora de evaluar el efecto global. El uso de pruebas basadas en la distribución normal para analizar datos que no tienen una distribución normal sí es hacer trampa.

Si las pruebas estadísticas usadas en el artículo son poco claras, ¿por qué las escogieron los autores, e incluyeron una referencia?

En ocasiones, parece que hay un número infinito de posibles pruebas estadísticas. De hecho, la mayoría de los estudios clínicos básicos se pueden analizar mediante alrededor de una docena. El resto son secundarias y deben reservarse para indicaciones especiales. Si el artículo que estamos leyendo parece describir un conjunto estándar de datos que se han recogido de una manera estándar, pero el nombre de la prueba utilizada es impronunciable y no aparece en un texto básico de estadística, puede que haya gato encerrado. En este caso, los autores deberían aclarar por qué han utilizado esta prueba e indicar una referencia (con números de página) donde se describa de forma detallada.

¿Se han analizado los datos de acuerdo con el protocolo original del estudio?

Aunque no estemos interesados en la justificación estadística, el sentido común debería indicarnos por qué los puntos 8 y 9 del cuadro 5.2 que aparece al final de este apartado son una forma muy grave de hacer trampa. Si se realiza una búsqueda lo suficientemente exhaustiva, inevitablemente se acabará por encontrar una categoría de participantes con unos resultados especialmente buenos o malos. Sin embargo, cada vez que evaluamos si un subgrupo en particular es diferente del resto se aumenta en gran medida la probabilidad de que al final se encuentre uno que parezca serlo, aunque la diferencia se deba por completo al azar.

Cuadro 5.2 Pruebas de causalidad (v. referencia bibliográfica 14)

1. ¿Existe evidencia de experimentos reales en seres humanos?
2. ¿Es la asociación fuerte?
3. ¿Es la asociación concordante entre los estudios?
4. ¿Es la relación temporal apropiada (es decir, la causa propuesta precede al efecto propuesto)?
5. ¿Hay un gradiente dosis-respuesta (es decir, una causa de mayor cuantía se sigue de un efecto mayor)?
6. ¿La asociación tiene sentido epidemiológico?
7. ¿La asociación tiene sentido biológico?
8. ¿Es la asociación específica?
9. ¿Es la asociación análoga a una asociación de causalidad previamente demostrada?

Del mismo modo, si jugamos a cara o cruz con alguien, con independencia de la desventaja que llevemos, llegará un momento en que tengamos un punto de ventaja. La mayoría de la gente estaría de acuerdo en que dejar de jugar en ese momento no sería juego limpio. Lo mismo sucede con la investigación. Si hacemos todo lo posible para (al final) obtener un resultado aparentemente positivo también puede resultar inevitable que nos estemos autoengañando sobre la verosimilitud del estudio. La finalización antes de tiempo de un ensayo sobre una intervención por razones éticas cuando los participantes del grupo lo están pasando especialmente mal es diferente y se describe en otra parte[8].

Volver de nuevo a los datos y repasarlos en busca de resultados «interesantes» (análisis de subgrupos retrospectivo o, más coloquialmente, dragado de datos) puede dar lugar a falsas conclusiones[9,10]. En un estudio inicial sobre el uso de la aspirina en la prevención del ictus en pacientes predispuestos, los resultados mostraron un efecto significativo en ambos sexos combinados y un análisis de subgrupos retrospectivo parecía mostrar que el efecto se limitaba a los varones[11]. Esta conclusión hizo que la aspirina no se administrase a las mujeres durante muchos años hasta que los resultados de otros estudios (incluido una gran metaanálisis[12]) mostraron que este efecto de subgrupos era espurio.

Este y otros ejemplos se exponen en un artículo de Oxman y Guyatt[13] titulado «A consumer's guide to subgroup analysis» («Guía del consumidor sobre análisis de subgrupos») donde se reproduce una lista de comprobación útil para decidir si las diferencias aparentes en las respuestas de subgrupos son reales.

Datos pareados, colas y valores atípicos

¿Se realizaron las pruebas pareadas con datos pareados?

Los estudiantes a menudo tienen dificultades para decidir si se debe utilizar una prueba estadística pareada o no pareada para analizar sus datos. Lo cierto es que esta decisión no encierra ningún misterio. Si se mide algo dos veces en cada participante (p. ej., la presión arterial en decúbito y en bipedestación), es probable que estemos interesados no sólo en la diferencia media de presión arterial entre la bipedestación y el decúbito en toda la muestra, sino en cuánto varía la presión arterial de cada individuo con la posición. En este caso, los datos son pareados porque cada medición inicial se empareja con una medición posterior.

En este ejemplo, el emparejamiento se realiza al realizar la medición en la misma persona en las dos ocasiones, pero hay otras posibilidades (p. ej., dos mediciones cualquiera de la ocupación de camas de la misma planta de un hospital). En estos casos, es probable que los dos conjuntos de valores presenten una correlación significativa (es probable que la presión arterial de una persona medida la semana próxima sea más parecida a su presión arterial de hace siete días que a la medida hace una semana en otra persona escogida al azar). Dicho de otro modo, es de esperar que dos valores «pareados» seleccionados al azar sean más parecidos entre sí que dos valores «no pareados» seleccionados al azar. Si esto no se tiene en cuenta, realizando las pruebas con muestras «pareadas» adecuadas,

es posible que al final tengamos una estimación sesgada de la significación de nuestros resultados.

¿Se ha realizado una prueba de dos colas siempre que el efecto de una intervención pudiese haber sido negativo?

El concepto de una prueba con colas siempre me hace visualizar imágenes de demonios o serpientes, lo que tal vez pueda deberse a mi aversión a la estadística. En realidad, el término *cola* se refiere a los extremos de la distribución (las zonas oscuras de la fig. 5.1). Supongamos que la gráfica representa la presión arterial diastólica de un grupo de personas de las cuales una muestra aleatoria va a recibir una dieta baja en sodio. Si esta dieta hiposódica tuviese un efecto hipotensivo importante, sería más probable que las mediciones posteriores de la presión arterial en estos participantes estuviesen en la «cola» izquierda de la gráfica. Por lo tanto, tendríamos que analizar los datos con pruebas estadísticas diseñadas para mostrar si unas mediciones inusualmente bajas en esta muestra de pacientes tendrían probabilidades de haberse producido por azar.

Sin embargo, habría que plantearse si, además de suponer que una dieta hiposódica podría reducir la presión arterial, también podría *elevarla.* Aunque haya razones fisiológicas válidas que puedan explicar el efecto en este ejemplo en particular, no es científicamente correcto asumir siempre que sabemos la dirección del efecto que tendrá nuestra intervención. Un nuevo fármaco destinado a disminuir las náuseas en realidad podría agravarlas y un folleto educativo dirigido a reducir la ansiedad podría aumentarla. Por lo tanto, el análisis estadístico debería, por lo general, probar la hipótesis de que los valores altos *o* bajos de nuestro conjunto de datos hayan surgido por azar. En el lenguaje estadístico, esto significa que se debe utilizar una prueba de dos colas a menos que tengamos una evidencia muy convincente de que la diferencia sólo puede ir en una dirección.

¿Se analizaron los valores atípicos con sentido común y se realizaron los ajustes estadísticos apropiados?

Unos resultados inesperados pueden reflejar características idiosincrásicas del participante (p. ej., un metabolismo inusual) o bien errores de medición (p. ej., equipo defectuoso), de interpretación (p. ej., lectura incorrecta de un contador) o de cálculo (p. ej., comas decimales mal situadas). Sólo el primero de estos ejemplos es un resultado «real» que debe incluirse en el análisis. Un resultado que esté varios órdenes de magnitud alejado de los demás tiene menos probabilidades de ser verdadero, pero podría serlo. Hace unos años, cuando estaba realizando un proyecto de investigación, medí varias concentraciones hormonales distintas en unos 30 participantes. La cifra de la hormona de crecimiento de uno de los participantes era unas cien veces superior a la de todos los demás. Supuse que se trataba de un error de transcripción, por lo que moví la coma decimal dos lugares a la izquierda. Varias semanas después, el técnico que había analizado las muestras me preguntó qué había pasado con el paciente que tenía acromegalia.

Desde el punto de vista estadístico, la corrección de los valores atípicos (p. ej., para modificar su efecto sobre el resultado global) es una maniobra estadística bastante sofisticada. Los lectores que estén interesados pueden consultar la sección correspondiente en su manual de estadística favorito.

Correlación, regresión y causalidad

¿Se ha distinguido la correlación de la regresión, y se ha calculado e interpretado adecuadamente el coeficiente de correlación («valor *r*»)?

Para muchas personas sin formación estadística, los términos *correlación* y *regresión* son sinónimos, y se refieren vagamente a una imagen mental de una gráfica de dispersión con puntos salpicados desordenadamente a lo largo de una línea diagonal que surge de la intersección de los ejes. Es acertado suponer que si dos cosas no están correlacionadas, no tendrá sentido intentar una regresión. Sin embargo, la regresión y la correlación son términos estadísticos precisos que tienen funciones diferentes[2].

El valor *r* (o según su denominación oficial, «coeficiente de correlación producto-momento de Pearson») es uno de los instrumentos estadísticos más usados en este libro. En sentido estricto, el valor *r* no es válido a menos que se cumplan ciertos criterios, los cuales se exponen a continuación:

1. Los datos (o, más precisamente, la población de la que se extraen los datos) deben tener una distribución normal. En caso contrario, se deben utilizar en su lugar pruebas no paramétricas de correlación (v. tabla 5.1).
2. Las dos variables deben ser estructuralmente independientes (es decir, una no debe variar obligatoriamente con la otra). Si no lo son, debe utilizarse en su lugar una *t* pareada u otra prueba pareada.
3. Sólo debe realizarse un par de mediciones en cada participante, pues las mediciones realizadas en participantes sucesivos deben ser estadísticamente independientes entre sí si se quiere obtener estimaciones no sesgadas de los parámetros de la población de interés.
4. Cada valor *r* debe acompañarse de un valor *p,* que expresa cuál es la probabilidad de que una asociación con esta fuerza haya surgido por azar (v. sección «¿Se han calculado e interpretado adecuadamente los valores *p*?») o de un intervalo de confianza, que expresa el rango dentro del cual es probable que se encuentre el verdadero valor *R* (v. sección «¿Se han calculado los intervalos de confianza, y se reflejan en las conclusiones de los autores?»). (Debe tenerse en cuenta que la «*r*» minúscula representa el coeficiente de correlación de la muestra, mientras que la «*R*» mayúscula representa el coeficiente de correlación de toda la población.)

Debe recordarse también que, aunque es apropiado calcular el valor *r* a partir de un conjunto de datos, no indicará si la relación es causal, por muy significativa que sea (v. más adelante).

El término *regresión* hace referencia a una *ecuación* matemática que permite predecir una variable (la variable de *destino*) a partir de otra (la variable *independiente*). Por lo tanto, la regresión implica una dirección de la influencia aunque, como se argumentará en la siguiente sección, no demuestra la causalidad. En una regresión múltiple, una ecuación matemática mucho más compleja (de la cual, por fortuna, se encarga el ordenador que la calcula) permite predecir la variable de destino a partir de dos o más variables independientes (a menudo denominadas *covariables*).

La ecuación de regresión más simple, que el lector tal vez recuerde de sus días de colegio, es $y = a + bx$, donde y es la variable dependiente (representada en el eje vertical), x es la variable independiente (representada en el eje horizontal), a es la ordenada en el origen (intersección con el eje y) y b es una constante. No hay muchas variables biológicas que se puedan predecir con una ecuación tan simple. El peso de un grupo de personas, por ejemplo, varía con su talla, pero no de forma lineal. En la primera edición de este libro, se ponía el siguiente ejemplo: soy el doble de alta que mi hijo y peso el triple, pero, aunque soy cuatro veces más alta que mi sobrino recién nacido, peso mucho más de seis veces que él. Tanto mi hijo como mi sobrino ahora me sacan una cabeza, pero el ejemplo sigue siendo válido. Es probable que el peso varíe más estrechamente en función del cuadrado de la talla de una persona que con la propia talla, por lo que sería más adecuado utilizar una regresión cuadrática en lugar de una de tipo lineal.

Aunque se hayan introducido suficientes datos de peso y talla en un ordenador para calcular la ecuación de regresión que mejor prediga el peso de una persona a partir de su talla, las predicciones seguirán siendo bastante imprecisas, pues el peso y la talla no presentan una *correlación* tan estrecha. Hay otros factores que influyen en el peso, además de la talla, y sería posible (para ilustrar el principio de regresión múltiple) introducir en el ordenador datos sobre la edad, el sexo, la ingesta diaria de calorías y el nivel de actividad física, y preguntarle en qué medida cada una de estas covariables contribuye a la ecuación (o modelo) global.

Los principios elementales descritos aquí, sobre todo los puntos numerados anteriormente, deberían ayudarnos a detectar si la correlación y la regresión se utilizan correctamente en el artículo que estamos leyendo. Una exposición más detallada sobre el tema se puede encontrar en los manuales estadísticos que figuran al final de este capítulo[5-7] y en el cuarto artículo de la serie *Basic Statistics for Clinicians*[2].

¿Se han hecho suposiciones sobre la naturaleza y la dirección de la causalidad?

Siempre hay que tener en cuenta la falacia ecológica: sólo porque una ciudad tenga un gran número de desempleados y una tasa de delincuencia muy alta, no quiere decir necesariamente que los desempleados estén cometiendo los delitos. Dicho de otro modo, la presencia de una *asociación* entre A y B no implica nada en absoluto sobre la presencia o la dirección de la causalidad. Para demostrar

que A ha *causado* B (en lugar de que B haya causando A, o que C haya causado A y B), se necesita algo más que un coeficiente de correlación. En el cuadro 5.1 se presentan varios criterios, desarrollados inicialmente por Sir Austin Bradford Hill[14], que deben cumplirse antes de asumir la causalidad.

Probabilidad y confianza

¿Se han calculado e interpretado adecuadamente los valores *p*?

Uno de los primeros valores que un estudiante de estadística aprende a calcular es el valor *p*, que es la probabilidad de que cualquier resultado en particular haya surgido por azar. En la práctica científica estándar, que es esencialmente arbitraria, se suele considerar que un valor *p* menor de 1/20 (expresado como $p < 0{,}05$, y equivalente a una apuesta de veinte a uno) es estadísticamente significativo, y un valor *p* de menor de 1/100 ($p < 0{,}01$) como estadísticamente muy significativo.

Por consiguiente, y por definición, una asociación al azar de cada veinte (alrededor de un resultado importante publicado en cada número de revista) parecerá ser significativa cuando no lo es, y una de cada cien parecerá muy significativa cuando en realidad no es más que un golpe de suerte. Por lo tanto, si los investigadores han hecho comparaciones múltiples, deberían realizar una corrección para tratar de tenerlo en cuenta. Es probable que el procedimiento más conocido para llevarlo a cabo sea la prueba de Bonferroni (que se describe en la mayoría de los manuales estándar de estadística), aunque un revisor de las ediciones previas de este libro consideró que esta prueba es «demasiado estricta» y propuso otras. En lugar de especular sobre pruebas que no comprendo personalmente, recomiendo que se pida consejo a un estadístico si en el artículo que se está leyendo se plantean comparaciones múltiples.

Un resultado en el rango estadísticamente significativo ($p < 0{,}05$ o $p < 0{,}01$ en función de lo que se haya elegido como punto de corte) sugiere que los autores deberían rechazar la hipótesis nula (es decir, la hipótesis de que no existe una diferencia real entre dos grupos). Sin embargo, como ya se ha expuesto antes (v. sección ¿Se abordaron las cuestiones estadísticas preliminares?), un valor *p* en el rango no significativo indica que *o bien* no hay ninguna diferencia entre los grupos *o bien* había demasiado pocos participantes para demostrar dicha diferencia en caso de que existiese, pero no aclara cuál de estas dos posibilidades es la responsable.

El valor *p* tiene una limitación adicional. Por lo tanto, Guyatt y cols., en el primer artículo de su serie *Basic Statistics for Clinicians* sobre la evaluación de hipótesis utilizando los valores *p* concluyen lo siguiente:

> *¿Por qué utilizar un único punto de corte (de significación estadística) cuando la elección de dicho punto es arbitraria? ¿Por qué convertir la pregunta de si un tratamiento es eficaz en una dicotomía (decisión de sí o no) cuando sería más apropiado considerarla un continuo?*[1].

Para ello, necesitamos intervalos de confianza, que se describen a continuación.

¿Se han calculado los intervalos de confianza, y se reflejan en las conclusiones de los autores?

Un intervalo de confianza, que un buen estadístico puede calcular a partir del resultado de casi cualquier prueba estadística (prueba *t,* valor *r,* reducción del riesgo absoluto [RRA], número necesario a tratar, así como sensibilidad, especificidad y otras características principales de una prueba diagnóstica) permite estimar, tanto para los ensayos «positivos» (los que muestran una diferencia estadísticamente significativa entre los dos grupos del ensayo) como para los «negativos» (los que parecen mostrar que no existe diferencia), si la evidencia es *fuerte* o *débil,* y si el estudio es *definitivo* (es decir, si evita la necesidad de realizar más estudios similares). El cálculo de los intervalos de confianza se ha descrito con gran claridad en el clásico *Statistics with confidence*[15], y Guyatt y cols. han escrito sobre su interpretación[4].

Si repitiéramos cientos de veces el mismo ensayo clínico, no obtendríamos exactamente el mismo resultado cada vez. Pero, *como promedio,* estableceríamos un nivel particular de diferencia (o de falta de diferencia) entre los dos grupos del ensayo. En el 90% de los ensayos, la diferencia entre los dos grupos se situaría dentro de ciertos límites amplios, y en el 95% de los ensayos, estaría entre ciertos límites aún más amplios.

Ahora bien, si, como suele suceder, sólo se realizase un ensayo, ¿cómo se puede saber lo cerca que está el resultado de la diferencia real entre los grupos? La respuesta es que no es posible saberlo, pero mediante el cálculo, por ejemplo, del intervalo de confianza del 95% alrededor de nuestro resultado, podremos afirmar que hay un 95% de posibilidades de que la diferencia «real» se encuentre entre esos dos límites. La afirmación que debe buscarse en un artículo debe decir algo como esto:

> *En un estudio sobre el tratamiento de la insuficiencia cardíaca, el 33% de los pacientes asignados de forma aleatoria para recibir IECA fallecieron, mientras que el 38% de los asignados de forma aleatoria para recibir hidralazina y nitratos fallecieron. La estimación puntual de la diferencia entre los grupos (la mejor estimación individual del beneficio de vidas salvadas gracias al uso de un IECA) es del 5%. El intervalo de confianza del 95% alrededor de esta diferencia es de −1,2% a +12%.*

Lo más probable es que los resultados se expresen del siguiente modo resumido:

> *La supervivencia del grupo del IECA era un 5% (IC del 95% [−1,2, +12]) mayor.*

En este ejemplo en particular, el intervalo de confianza del 95% se superpone con una diferencia cero y, si estábamos expresando el resultado como una

dicotomía (es decir, la hipótesis se «demuestra» o se «rechaza»), lo clasificaríamos como un ensayo negativo. Sin embargo, como alegan Guyatt y cols., *probablemente* exista una diferencia real y *probablemente* se encuentre más cerca del 5% que del 1,2% o del +12%. Una conclusión más útil a partir de estos resultados es que «a igualdad de todos los demás factores, un inhibidor de la enzima convertidora de angiotensina (IECA) probablemente sea la opción apropiada para los pacientes con insuficiencia cardíaca, pero la fuerza de esta inferencia es débil»[4].

Como se indica en la sección «Diez preguntas que deben plantearse sobre un artículo que pretende validar una prueba diagnóstica o de cribado», cuanto mayor será el ensayo (o mayores los resultados combinados de varios ensayos), más estrecho será el intervalo de confianza y, por lo tanto, más probable será que el resultado sea definitivo.

A la hora de interpretar los ensayos «negativos», es importante saber si «sería probable que un ensayo mucho más grande mostrase un beneficio significativo». Para responder a esta pregunta, hay que observar el límite *superior* del intervalo de confianza del 95% del resultado. Sólo hay una posibilidad entre cuarenta (es decir, una probabilidad del 2,5%, pues el otro 2,5% de resultados extremos estará por debajo del límite *inferior* del intervalo de confianza del 95%) de que el resultado real sea igual o mayor que esta cifra. Ahora nos debemos preguntar si este nivel de diferencia sería *clínicamente* significativo y, si no lo fuera, el ensayo ya se puede clasificar no sólo como negativo sino también como definitivo. Además, si el límite superior del intervalo de confianza del 95% representase un nivel de diferencia clínicamente significativo entre los grupos, el ensayo puede ser negativo, pero también es no definitivo.

Hasta hace poco, el uso de intervalos de confianza era relativamente infrecuente en los artículos médicos. Por fortuna, la mayoría de los ensayos publicados en revistas que siguen las directrices CONSORT (Consolidated Standards of Reporting Trials, o Normas consolidadas para la publicación de ensayos clínicos) (v. sección «Ensayos controlados aleatorizados») ahora los incluyen de forma sistemática, pero aun así, muchos autores no interpretan sus intervalos de confianza correctamente. Se debe comprobar cuidadosamente la sección de discusión para ver si los autores han concluido correctamente (i) si su ensayo apoya su hipótesis y en qué medida, y (ii) la necesidad de llevar a cabo ensayos adicionales.

Resultado final

¿Han expresado los autores los efectos de una intervención en términos del beneficio o perjuicio probable que puede esperar un paciente individual?

Está muy bien afirmar que una intervención particular produce una «diferencia estadísticamente significativa» en el resultado, pero si nos pidiesen que tomásemos un fármaco querríamos saber cuánto mejorarían nuestras probabilidades (en términos de cualquier resultado en particular) que si no lo tomásemos. Hay tres cálculos sencillos (que de verdad *son* simples: si se sabe sumar, restar, multiplicar

y dividir, será posible comprender esta sección) que nos permitirán responder a esta pregunta con objetividad y de manera que signifique algo para personas sin experiencia en estadística. Los cálculos son la reducción del riesgo relativo, la RRA y el número necesario a tratar.

Para ilustrar estos conceptos y convencer al lector de que es necesario conocerlos, tomaremos como ejemplo un estudio que Fahey y cols.[16] realizaron hace unos años. Estos autores escribieron a 182 miembros de un organismo sanitario gubernamental de Inglaterra (que eran responsables en alguna medida de tomar decisiones importantes sobre los servicios sanitarios) y les expusieron los siguientes datos acerca de cuatro programas de rehabilitación diferentes para víctimas de infartos de miocardio. Les preguntaron cuál preferirían financiar:

Programa A: reducía la mortalidad un 20%.

Programa B: producía una reducción absoluta de fallecimientos del 3%.

Programa C: aumentaba la tasa de supervivencia de los pacientes del 84% al 87%.

Programa D: requería la inclusión de 31 personas en el programa para evitar un fallecimiento.

De los 140 miembros del organismo sanitario que respondieron, sólo tres advirtieron que los cuatro «programas» se relacionaban en realidad con la misma serie de resultados. Los otros 137 participantes escogieron uno de los programas preferentemente respecto a los demás, lo que revela (además de su propia ignorancia) la necesidad de mejorar la formación básica en epidemiología de las autoridades sanitarias. En realidad, el Programa A es la reducción del riesgo relativo, el Programa B es la RRA, el Programa C es otra manera de expresar la RRA y el Programa D es el número necesario a tratar.

Continuando con este ejemplo, que Fahey y cols. tomaron de un estudio realizado por Yusuf y cols.[17], las cifras se pueden presentar en una tabla de dos por dos donde se ofrecen los detalles del tratamiento que recibieron los pacientes en su ensayo aleatorizado, y de si estaban vivos o muertos 10 años después (tabla 5.2).

Unos cálculos simples nos muestran que los pacientes que reciben tratamiento médico tienen una probabilidad de 404/1.325 = 0,305 o 30,5% de estar muertos a los 10 años. Este es el *riesgo absoluto* de fallecer del grupo control (tratamiento médico) y lo llamaremos *x*. Los pacientes asignados al azar al grupo de cirugía de revascularización coronaria tienen una probabilidad de 350/1.324 = 0,264 o 26,4%

Tabla 5.2 Datos de un ensayo comparativo entre un tratamiento médico y un injerto de revascularización coronaria tras un infarto de miocardio[16,17]

Tratamiento	Resultado a los 10 años (muertos/vivos)	Número total de pacientes asignados de forma aleatoria a cada grupo
Tratamiento médico	404/921	1.325
Revascularización coronaria	350/974	1.324

de estar muertos a los 10 años. Éste es el riesgo absoluto de fallecer del grupo de intervención (revascularización coronaria) y lo llamaremos *y*.

El *riesgo relativo* de fallecer de los pacientes de revascularización coronaria en comparación con los controles de intervención médica es y/x, es decir, 0,264/0,305 = 0,87 (87%).

La *reducción del riesgo relativo* (es decir, la cantidad que se redujo el riesgo de fallecer en el grupo de revascularización coronaria en comparación con el grupo control) es de $100 - 87\%$ $(1 - y/x) = 13\%$.

La *RRA* (o diferencia de riesgo), es decir, la cantidad absoluta que la revascularización coronaria reduce el riesgo de fallecer a los 10 años es del 30,5 − 26,4% = 4,1% (0,041).

El *número necesario a tratar*, es decir, el número de pacientes que deben someterse a revascularización coronaria para evitar, como promedio, un fallecimiento a los 10 años es el recíproco de la RRA: 1/RRA = 1/0,041 = 24.

Las fórmulas generales para calcular estos efectos «finales» de una intervención se reproducen en el apéndice 2. En el artículo de Jaeschke y cols. publicado en la serie Basic Statistics for Clinicians se explica cuál de estos valores es más útil en cada circunstancia[3].

Resumen

Si confiamos en la competencia estadística (y/o la honestidad intelectual) de los autores de un artículo, podemos ser víctimas de un grave engaño. La estadística puede ser una ciencia intimidante y la comprensión de sus puntos más sutiles a menudo requiere la ayuda de expertos. Sin embargo, espero que este capítulo haya demostrado que la estadística utilizada en la mayoría de los artículos de investigación médica puede evaluarse (al menos hasta cierto punto) por parte de personas no expertas usando una simple lista de comprobación como la que aparece en el apéndice 1. Además, aconsejo al lector que verifique los artículos que vaya a leer (o a escribir) para ver si incurren en los errores frecuentes indicados en el cuadro 5.2.

Bibliografía

1 Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ: Canadian Medical Association Journal 1995;**152**(1):27.
2 Guyatt G, Walter S, Shannon H, et al. Basic statistics for clinicians: 4. Correlation and regression. CMAJ: Canadian Medical Association Journal 1995;**152**(4):497.
3 Jaeschke R, Guyatt G, Shannon H, et al. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. CMAJ: CanadianMedical Association Journal 1995;**152**(3):351.
4 Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. CMAJ: Canadian Medical Association Journal 1995;**152**(2):169.
5 Norman GR, Streiner DL. *Biostatistics: the bare essentials*. USA: PMPH-USA; 2007.

6 Bowers D. *Medical statistics from scratch: an introduction for health professionals*. Oxford: JohnWiley & Sons, 2008.
7 Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 2000.
8 Pocock SJ. When (not) to stop a clinical trial for benefit. JAMA: The Journal of the American Medical Association 2005;**294**(17):2228-30.
9 Cuff A. Sources of Bias in Clinical Trials. 2013. http://applyingcriticality.wordpress.com/2013/06/19/sources-of-bias-in-clinical-trials/(accessed 26th June 2013).
10 Delgado-Rodríguez M, Llorca J. Bias. Journal of Epidemiology and Community Health 2004;**58**(8):635-41 doi: 10.1136/jech.2003.008466.
11 Group CCS. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. The New England Journal of Medicine 1978;**299**(2):53-9.
12 Antiplatelet Trialists' Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. British Medical Journal (Clinical Research Edition) 1988;**296**(6618):320.
13 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Annals of Internal Medicine 1992;**116**(1):78-84.
14 Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine 1965;**58**(5):295.
15 Altman DG, Machin D, Bryant TN, et al. *Statisticswith confidence: confidence intervals and statistical guidelines*. London: BMJ Books; 2000.
16 Fahey T, Griffiths S, Peters T. Evidence based purchasing: understanding results of clinical trials and systematic reviews. BMJ: British Medical Journal 1995;**311**(7012):1056-9.
17 Yusuf S, Zucker D, Passamani E, et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. The Lancet 1994;**344**(8922):563-70.

Capítulo 6 Artículos que describen ensayos de tratamientos farmacológicos y otras intervenciones sencillas

«Evidencia» y marketing

En este capítulo se describe la evaluación de la evidencia de los ensayos clínicos y la mayor parte de esta evidencia se refiere a los fármacos. Los médicos clínicos, enfermeras o farmacéuticos (es decir, quienes prescriben o dispensan fármacos) son profesionales que interesan a la industria farmacéutica, que gasta una parte de su presupuesto multimillonario anual de publicidad para tratar de influir en ellos (v. cuadro 6.1)[1]. En la actualidad, la industria puede ahora dirigir sus mensajes directamente a los pacientes a través de la publicidad directa al consumidor (PDC)[2]. Cuando escribí la primera edición de este libro en 1995, el tratamiento estándar de la candidiasis vaginal (infección por *Candida*) consistía en que el médico prescribiese óvulos de clotrimazol. Cuando se publicó la segunda edición en 2001, estos pesarios estaban disponibles sin receta en las farmacias. Durante los últimos 10 años, el clotrimazol se ha anunciado en el horario de máxima audiencia en televisión (por fortuna después de las nueve de la noche) y, más recientemente, los fabricantes de este medicamento y de otros fármacos potentes se publicitan a través de internet y de las redes sociales[3]. Por si alguien lo dudaba, este tipo de publicidad sutil tiende a hacer más hincapié en los beneficios que en los riesgos[4].

La manera más eficaz de cambiar los hábitos de prescripción de un médico es mediante un representante personal (conocido por muchos de nosotros como «visitador médico»), que visita en persona a los médicos con un maletín lleno de «evidencia» que respalda los beneficios de sus productos[5]. De hecho, como se comenta con más detalle en los capítulos 14 y 15, el movimiento de la medicina basada en la evidencia ha aprendido mucho de la industria farmacéutica en los últimos años sobre cómo modificar la conducta de los médicos y ahora utiliza las mismas técnicas sofisticadas de persuasión de la denominada *visita académica* de los profesionales sanitarios individuales[6]. Hay que destacar que la PDC suele ejercer su acción aprovechando el poder de persuasión del paciente, que se convierte en realidad en un «representante» no remunerado de la industria farmacéutica. Quien piense que es más fácil resistir la presión de un paciente que la de un representante real probablemente esté equivocado (en un ensayo controlado aleatorizado se demostró que la presión del paciente tiene un efecto

Cuadro 6.1 Diez consejos para la industria farmacéutica: cómo presentar su producto de forma idónea

1. Piense en un mecanismo fisiológico plausible de la acción del fármaco y ponga todas sus habilidades a la hora de presentarlo. Preferiblemente, encuentre un criterio de valoración indirecto que se vea fuertemente influenciado por el fármaco aunque pueda no ser estrictamente válido (v. sección «Toma de decisiones sobre el tratamiento»).
2. A la hora de diseñar ensayos clínicos, seleccione una población de pacientes, características clínicas y duración del ensayo que reflejen la respuesta máxima posible al fármaco.
3. Si es posible, compare su producto sólo con placebos. Si debe compararlo con un competidor, asegúrese de que éste se administra en una dosis subterapéutica.
4. Incluya los resultados de los estudios piloto en las cifras de los estudios definitivos de modo que parezca que el número de pacientes asignados de forma aleatoria es mayor que el número real.
5. No mencione ningún ensayo en el que haya habido algún fallecimiento o reacción adversa grave al fármaco en el grupo de tratamiento. Si es posible, no publique esos estudios.
6. Haga que el departamento gráfico maximice el impacto visual de su mensaje. Sirve de ayuda no poner indicaciones en los ejes de las gráficas y no aclarar si las escalas son lineales o logarítmicas. Asegúrese de no mostrar datos de pacientes individuales ni intervalos de confianza.
7. Conviértase en un maestro de las comparaciones sin comparador («mejor», ¿pero mejor que qué?).
8. Invierta la jerarquía estándar de la evidencia de modo que los casos aislados tengan prioridad sobre los ensayos aleatorizados y los metaanálisis.
9. Nombre al menos tres líderes de opinión locales que usen el fármaco y proporcione «muestras» para que el médico las pruebe.
10. Presente un «análisis de rentabilidad» que demuestre que su producto, aunque sea más caro que su competidor, «realmente sale más barato» (v. sección «El gran debate sobre las guías»).

significativamente mayor sobre la prescripción de los médicos después de haber recibido PDC sobre antidepresivos)[7].

Antes de aceptar un encuentro con un representante (o con un paciente armado con información obtenida de un artículo de periódico o de un página web de PDC), recuerde ciertas reglas básicas sobre el diseño de la investigación. Como se expuso en las secciones «Estudios de cohortes» y «Estudios transversales», las preguntas sobre los beneficios del tratamiento deberían resolverse idealmente con ensayos controlados aleatorizados. Sin embargo, las preguntas preliminares

sobre la farmacocinética (es decir, el comportamiento del fármaco hasta llegar a su sitio de acción), en especial las relativas a la biodisponibilidad, requieren un experimento sencillo sobre la posología en voluntarios sanos (y, si es ético y viable, en enfermos).

Se pueden recopilar las reacciones adversas frecuentes (y probablemente triviales) a un fármaco, y cuantificar su incidencia, en los ensayos controlados aleatorizados realizados para demostrar la eficacia de dicho fármaco. Sin embargo, las reacciones farmacológicas adversas infrecuentes (y por lo general más graves) requieren estudios de farmacovigilancia (recopilación prospectiva de datos sobre los pacientes que reciben un fármaco recién autorizado) y estudios de casos y controles (v. sección «Estudios de cohortes») para establecer la asociación. Lo ideal sería realizar experimentos de reexposición individuales (en los que el paciente que ha tenido una reacción atribuida al fármaco lo recibiese de nuevo en circunstancias supervisadas cuidadosamente) para establecer la causalidad[8].

Los representantes farmacéuticos ya no dicen tantas mentiras como solían (el marketing de los fármacos se ha convertido en una ciencia más sofisticada) aunque, como Goldacre[9] ha demostrado en su libro *Bad Pharma*, todavía proporcionan información que, en el mejor de los casos, es selectiva y, en el peor, claramente sesgada. A menudo, a los representantes les resulta útil para promocionar sus productos, por ejemplo, presentar resultados de ensayos no controlados y expresarlos en términos de diferencias antes/después de una medida de resultado particular. La lectura de la sección «Estudios transversales» y de la literatura sobre el efecto placebo[10,11] debería hacernos recordar por qué los estudios no controlados de tipo antes/después se publican en revistas para adolescentes, pero no en publicaciones científicas serias.

El Dr. Herxheimer, que fue editor del *Drug and Therapeutics Bulletin* durante muchos años, realizó un estudio sobre las «referencias» citadas en la publicidad de productos farmacéuticos en las principales revistas médicas de Reino Unido. Según me comentó, una proporción elevada de dichas referencias citan «datos de archivo» y otras muchas corresponden a publicaciones escritas, editadas y publicadas totalmente por la industria. Se ha demostrado que la evidencia de esas fuentes a veces (aunque no siempre) es de menor calidad científica que la que aparece en revistas independientes revisadas por expertos. Además, seamos realistas, si trabajásemos para una compañía farmacéutica que hubiese logrado un gran avance científico, probablemente remitiríamos nuestras conclusiones a una revista como *The Lancet* o *The New England Journal of Medicine* antes de presentarlo en una publicación propia. Dicho de otro modo, no hay por qué tirar a la basura los artículos sobre ensayos de fármacos *en función* de dónde se hayan publicado, pero debemos analizar detalladamente los métodos y análisis estadísticos de estos ensayos.

Toma de decisiones sobre el tratamiento

Sackett y cols.[8], en su libro *Clinical epidemiology – a basic science for clinical medicine,* sostienen que antes de administrar un fármaco a un paciente, el médico debe:

(a) Identificar el objetivo final del tratamiento *para este paciente* (curación, prevención de la recidiva, limitación de la discapacidad funcional, prevención de complicaciones posteriores, tranquilidad, paliación, alivio sintomático, etc.).
(b) Seleccionar el tratamiento *más adecuado* utilizando toda la evidencia disponible (esto incluye plantearse si el paciente tiene que tomar cualquier fármaco).
(c) Especificar la *meta del tratamiento* (¿cómo sabremos cuándo interrumpir el tratamiento, modificarlo o cambiar a algún otro?).

Por ejemplo, en el tratamiento de la hipertensión arterial, el médico puede decidir que:

(a) El *objetivo final del tratamiento* es evitar la lesión (adicional) de órganos diana (cerebro, ojos, corazón, riñones, etc.) y de ese modo prevenir el fallecimiento.
(b) La *elección de un tratamiento específico* se realiza entre las distintas clases de fármacos antihipertensivos seleccionados basándose en ensayos aleatorizados, controlados con placebo y comparativos, así como entre tratamientos no farmacológicos como la restricción de sal.
(c) La *meta del tratamiento* podría ser una presión arterial diastólica de fase V (brazo derecho o sedestación) menor de 90 mmHg, o lo más cercana posible a esa cifra que fuese tolerable frente a los efectos secundarios de los fármacos.

Si no se siguen estos tres pasos (como suele suceder, p. ej., en cuidados paliativos), puede producirse un caos terapéutico. En una crítica velada a los criterios de valoración indirectos, Sackett y cols. nos recuerdan que la elección del tratamiento específico debe basarse en la evidencia de lo que *funciona* y no en lo que *parece funcionar* o *debería funcionar.* Estos autores advierten de que «el tratamiento de hoy, cuando se basa en hechos biológicos o en una experiencia clínica no controlada, puede convertirse en la broma de mal gusto de mañana»[8].

Criterios de valoración indirectos

No he incluido esta sección tan sólo por mera afición. Para aquellos que sean clínicos en ejercicio (y no académicos), es posible que el contacto principal con los artículos publicados proceda de la literatura que ofrecen los «representantes farmacéuticos». La industria farmacéutica se mueve como pez en el agua a la hora de utilizar los criterios de valoración indirectos y no tengo ningún reparo en hacer hincapié en la necesidad de evaluar con sumo cuidado este tipo de medidas de resultado.

Definiré un criterio de valoración indirecto como «una variable que se mide con relativa facilidad y que predice un resultado poco frecuente o distante de un estímulo tóxico (p. ej., contaminante) o de una intervención terapéutica (p. ej., fármaco, procedimiento quirúrgico o consejo), pero que es no es en sí misma una medida directa de un perjuicio o de un beneficio clínico». El interés creciente en los criterios de valoración indirectos en la investigación médica refleja dos características importantes de su uso:

- Pueden reducir considerablemente el *tamaño muestral,* la *duración* y, por lo tanto, *el coste* de los ensayos clínicos.
- Pueden permitir evaluar los tratamientos en situaciones donde el uso de los resultados primarios sería excesivamente *invasivo* o *contrario a la ética.*

Cuando se evalúan los productos farmacéuticos, los criterios de valoración indirectos que suelen usarse son:

- Mediciones farmacocinéticas (p. ej., curvas de concentración/tiempo de un fármaco o de su metabolito activo en la sangre).
- Medidas *in vitro* (es decir, de laboratorio), como la concentración media inhibitoria de un antimicrobiano frente a un cultivo bacteriano en agar.
- Aspecto macroscópico de los tejidos (p. ej., imagen endoscópica de erosión gástrica).
- Variación de la concentración de (supuestos) «marcadores biológicos de la enfermedad» (p. ej., la microalbuminuria en la evaluación de la nefropatía diabética).
- Aspecto radiológico (p. ej., una veladura en una radiografía de tórax o, en un contexto más moderno, la resonancia magnética funcional).

Los criterios de valoración indirectos tienen una serie de inconvenientes. En primer lugar, una variación de este tipo de criterio no responde por sí misma a las preguntas preliminares esenciales: «¿cuál es el objetivo del tratamiento en este paciente?» y «¿cuál es, según los estudios de investigación válidos y fiables, el mejor tratamiento disponible para esta enfermedad?». En segundo lugar, es posible que el criterio de valoración indirecto no refleje con precisión cuál es la meta del tratamiento; dicho de otro modo, quizá no sea válido o fiable. En tercer lugar, el uso de un criterio de valoración indirecto tiene las mismas limitaciones que el uso de cualquier otra medida *individual* del éxito o fracaso de un tratamiento: no tiene en cuenta todas las demás medidas. Una confianza excesiva en un único criterio de valoración indirecto como medida del éxito terapéutico suele reflejar una perspectiva clínica limitada o ingenua.

Por último, los criterios de valoración indirectos suelen desarrollarse en modelos animales de la enfermedad ya que los cambios en una variable específica se pueden medir en condiciones controladas en una población bien definida. Sin embargo, la extrapolación de estos resultados a las enfermedades humanas puede ser inválida[12]:

- En estudios con animales, la población estudiada tiene unas características biológicas bastante uniformes y puede ser genéticamente endogámica.
- Tanto el tejido como la enfermedad que se está estudiando puede tener características importantes (p. ej., susceptibilidad al patógeno o velocidad de replicación celular) distintas a la enfermedad equivalente en el ser humano.
- Los animales se mantienen en un ambiente controlado, lo que minimiza la influencia de variables de estilo de vida (p. ej., dieta, ejercicio o estrés) y la medicación concomitante.
- La administración de dosis altas de sustancias químicas a animales de experimentación puede distorsionar las vías metabólicas habituales, lo que puede

dar lugar a resultados engañosos. Las especies animales más adecuadas para sustituir a los seres humanos varían según las diferentes sustancias químicas.

Las características ideales de un criterio de valoración indirecto se muestran en el cuadro 6.2. Si el representante que está tratando de convencernos de la utilidad del fármaco no puede justificar los criterios de valoración utilizados, habría que pedirle que ofreciese una evidencia adicional.

A continuación, se presentan algunos ejemplos reales de criterios de valoración indirectos que han provocado prácticas y recomendaciones erróneas:

- El uso de los hallazgos del ECG en lugar de los resultados clínicos (síncope, muerte) a la hora de evaluar la eficacia y la seguridad de los fármacos antiarrítmicos[13].

Cuadro 6.2 Características ideales de un criterio de valoración indirecto

1. El criterio de valoración indirecto debe ser fiable, reproducible, clínicamente disponible, fácil de cuantificar, asequible y presentar un efecto «dosis-respuesta» (es decir, cuanto mayor sea el nivel del criterio de valoración indirecto, mayor será la probabilidad de presentar la enfermedad).
2. Debe ser un auténtico factor predictivo de la enfermedad (o del riesgo de la enfermedad) y no limitarse a expresar la exposición a una covariable. La relación entre el criterio de valoración indirecto y la enfermedad debe tener una explicación plausible desde el punto de vista biológico.
3. Debe ser sensible, es decir, un resultado «positivo» del criterio de valoración indirecto debe detectar a todos o a la mayoría de los pacientes con un riesgo mayor de resultados adversos.
4. Debe ser específico, es decir, un resultado «negativo» debe excluir a todas o a la mayoría de las personas que no tengan un riesgo mayor de resultado adverso.
5. Debe haber un punto de corte preciso entre los valores normales y anormales.
6. Debe tener un valor predictivo positivo aceptable, es decir, un resultado «positivo» siempre o casi siempre debe significar que el paciente identificado con dicho criterio tiene mayor riesgo de resultados adversos (v. sección «Diez preguntas que deben plantearse sobre un artículo que describa una intervención compleja»).
7. Debe tener un valor predictivo negativo aceptable, es decir, un resultado «negativo» siempre o casi siempre debe significar que el paciente identificado con dicho criterio no tiene un riesgo mayor de resultados adversos (v. sección «Diez preguntas que deben plantearse sobre un artículo que describa una intervención compleja»).
8. Debe ser posible someterlo a un control de calidad.
9. Los cambios del criterio de valoración indirecto deben reflejar con rapidez y precisión la respuesta al tratamiento, sobre todo los niveles deben normalizarse en las fases de remisión o curación.

- El uso de los hallazgos de las radiografías en lugar de los resultados clínicos (dolor, pérdida de función) para monitorizar la progresión de la artrosis y la eficacia de los fármacos modificadores de la enfermedad[14].
- El uso de la albuminuria en lugar del balance global beneficio clínico/perjuicio para evaluar la utilidad de bloqueo dual renina-angiotensina en la hipertensión[15,16]. En este ejemplo, la intervención se basó en el argumento hipotético según el cual el bloqueo de la vía renina-angiotensina en dos etapas distintas sería el doble de eficaz y el criterio de valoración indirecto confirmó que esto parecía ser cierto, pero la combinación también fue el doble de eficaz a la hora de producir hipopotasemia, un efecto secundario potencialmente mortal.

Sería injusto sugerir que la industria farmacéutica siempre utiliza los criterios de valoración indirectos con la intención deliberada de engañar a las autoridades encargadas de autorizar los fármacos y a los profesionales sanitarios. Como ya se ha comentado en la sección «Evidencia y marketing», los criterios de valoración indirectos deben responder a criterios éticos y económicos. Sin embargo, la industria tiene gran interés en exagerar los argumentos que respaldan estos criterios de valoración[9], por lo que hay que ser cauteloso al leer un artículo cuyas conclusiones no se basen en «resultados objetivos relevantes para los pacientes».

Los criterios de valoración indirectos son sólo una de las muchas formas en las que los ensayos patrocinados por la industria pueden dar una impresión engañosa de la eficacia de un fármaco. Otras influencias sutiles (y no tan sutiles) sobre el diseño de la investigación, como plantear la pregunta de una manera especial o la publicación selectiva de los resultados, se han descrito en una revisión Cochrane reciente que explica cómo los ensayos patrocinados por la industria tienden a favorecer sus propios productos[17].

¿Qué información debería proporcionar un artículo que describa un ensayo controlado aleatorizado?: declaración CONSORT

Los ensayos sobre fármacos son un ejemplo de una «intervención sencilla»: una intervención que está bien delimitada (es decir, es fácil describir en qué consiste la intervención) y que se presta a un diseño de investigación que compara un «grupo con intervención» frente a un «grupo sin intervención». En los capítulos 3 y 4 se ofrecieron varios consejos preliminares sobre la evaluación de la calidad metodológica de los estudios de investigación. Ahora se ahondará en ello con más detalle. En 1996, un grupo de trabajo internacional elaboró una lista de comprobación estándar, denominada CONSORT (Consolidated Standards of Reporting Trials, Normas consolidadas para la publicación de ensayos clínicos), para la publicación de ensayos controlados aleatorizados en revistas médicas. Esta lista se ha actualizado varias veces, la última en 2010[18]. Sin duda, el uso de estas listas de comprobación ha aumentado la calidad y homogeneidad de las publicaciones de ensayos en la literatura médica[19]. En la tabla 6.1 se muestra una lista de comprobación basada en la declaración CONSORT. No es necesario

Tabla 6.1 Lista de comprobación para un ensayo controlado aleatorizado basada en la declaración CONSORT (v. referencia 17).

Título/resumen	¿Se indica en el título y en el resumen cómo se asignaron los participantes a las intervenciones (p. ej., «asignación aleatoria», «aleatorizado» o «asignado de forma aleatoria»)?
Introducción	¿Se han explicado adecuadamente los antecedentes y fundamentos científicos para el estudio?
Métodos	
Objetivos	¿Se han descrito explícitamente los objetivos específicos y/o hipótesis que se va a probar?
Participantes y contexto	¿Se indican en el artículo los criterios de selección de los participantes y los contextos y lugares donde se recogieron los datos?
Intervenciones	¿El artículo proporciona detalles precisos de la intervención(es) y de la intervención(es) de control y de cuándo se administraron?
Resultados	¿Se han definido claramente las variables de resultado principales y secundarias? Si corresponde, ¿se han dispuesto los métodos usados para aumentar la calidad de las mediciones (p. ej., observaciones múltiples, entrenamiento de los asesores)?
Tamaño muestral	¿Cómo se determinó el tamaño muestral? Si corresponde, ¿se han explicado y justificado los análisis intermedios y/o las reglas para interrumpir prematuramente el estudio?
Estudio ciego (enmascaramiento)	¿Se indica en el artículo si los participantes, quienes administraron las intervenciones y quienes evaluaron los resultados desconocían la asignación de los grupos? ¿Cómo se evaluó si el enmascaramiento fue correcto?
Métodos estadísticos	¿Fueron apropiados los métodos estadísticos utilizados para comparar los grupos según los resultados principales y secundarios, y cualquier análisis de subgrupos?
Detalles de la asignación aleatoria	
Generación de secuencia	¿Se describió con claridad el método usado para generar la secuencia de asignación aleatoria, incluidos los detalles de cualquier restricción (bloques, estratificación)?
Ocultación de la asignación	¿Se describió el método utilizado para implementar la secuencia de asignación aleatoria (p. ej., recipientes numerados o teléfono central) y se dejó claro si la secuencia permaneció oculta hasta la asignación de las intervenciones?
Implementación	¿Se indica en el artículo quién generó la secuencia de asignación, quién reclutó a los participantes y quién asignó a los participantes a sus grupos?

(Continúa)

Tabla 6.1 ***(cont.)***

Resultados	
Diagrama de flujo	¿Se incluye un diagrama claro que muestre el flujo de los participantes durante el ensayo? Este diagrama debería indicar, para cada grupo, los números de pacientes asignados al azar, los que recibieron el tratamiento propuesto, los que completaron el protocolo del estudio y los que se analizaron según el resultado principal
Desviaciones del protocolo	¿Se han explicado y justificado todas las desviaciones del protocolo original del estudio?
Datos de reclutamiento	¿Los autores han indicado el rango de fechas de reclutamiento de los participantes en el estudio?
Datos iniciales	¿Se describen las características demográficas y clínicas iniciales de cada grupo?
Números analizados	¿Se incluye el número de participantes (denominador) de cada grupo en cada análisis, y el análisis es de tipo «por intención de tratar»?
Resultados y estimación	Para cada resultado principal y secundario, ¿hay un resumen de resultados para cada grupo, así como la magnitud del efecto estimado y su precisión (p. ej., intervalo de confianza del 95%)?
Análisis secundarios	¿Se han descrito y justificado todos los análisis adicionales, incluidos los análisis de subgrupos, tanto especificados a priori como exploratorios?
Efectos adversos	¿Han descrito y comentado los autores todos los efectos adversos importantes?
Discusión	
Interpretación	¿Está justificada la interpretación de los resultados, teniendo en cuenta las hipótesis del estudio, las fuentes de posibles sesgos o de imprecisión y los peligros de comparaciones múltiples?
Generalización	¿Han realizado los autores una estimación justificable de la generalización (validez externa) de los resultados del ensayo?

aprenderse esta tabla de memoria, pero hay que recurrir a ella a la hora de evaluar críticamente un artículo en el que pueda utilizarse o si estamos planeando realizar un ensayo aleatorizado.

A propósito, una forma importante de reducir el sesgo en el marketing de los fármacos es garantizar que todos los ensayos que se *comiencen* también se *escriban* y se *publiquen*[20]. De lo contrario, la industria farmacéutica (o cualquiera con un interés personal) podría impedir la publicación de todos los ensayos que no respalden su propia creencia en la eficacia y/o rentabilidad de un producto en especial. Goldacre[9] ha tratado en su libro el tema del registro obligatorio al comenzar los ensayos (y la reticencia de algunas compañías farmacéuticas a la hora de aceptarlo).

Cómo obtener evidencia útil de un representante farmacéutico

Cualquier médico que haya hablado con un visitador que quiera presentar un fármaco antiinflamatorio no esteroideo estará familiarizado con el ejemplo de la erosión gástrica. La pregunta que debemos plantearle no es: «¿cuál es la incidencia de erosión gástrica de su fármaco?», sino: «¿cuál es la incidencia de hemorragia gástrica potencialmente mortal?». A continuación se recogen varias preguntas adicionales que deben plantearse a los representantes farmacéuticos, según un artículo publicado en la revista *Drug and Therapeutics Bulletin*[21]. Pueden consultarse consejos más detallados sobre cómo desmontar los argumentos de los ensayos clínicos patrocinados que tratan de ocultarnos la realidad con estadísticas en la *Users' Guide* de Montori y cols.[22] y (de un modo más tangencial, pero que merece citarse) en el libro superventas de Goldacre sobre los trucos corporativos de las grandes empresas farmacéuticas[9]:

1. Vea a los representantes sólo con cita previa. Escoja hablar sólo con aquellos cuyo producto resulte de interés y limite la entrevista a ese producto.
2. Tome las riendas de la entrevista. No se limite a escuchar una charla comercial rutinaria ensayada sino que pregunte directamente por la información.
3. Solicite evidencia independiente publicada en revistas acreditadas y con revisión experta.
4. Ignore los folletos promocionales, pues a menudo contienen material no publicado, gráficos engañosos y citas selectivas.
5. Ignore la «evidencia» anecdótica, como la prescripción del producto por parte de un médico famoso.
6. Utilice el acrónimo STEP para solicitar evidencia en cuatro áreas específicas:
   - *S*eguridad, es decir, probabilidad de efectos secundarios a largo plazo o graves causados por el fármaco (debe recordarse que las reacciones adversas infrecuentes, pero graves, a los nuevos fármacos pueden estar mal documentadas).
   - *T*olerabilidad, que se mide mejor comparando las tasas de abandono acumuladas del fármaco y de su competidor más significativo.
   - *E*ficacia, cuya dimensión más relevante es la eficacia comparada entre el producto estudiado y nuestro fármaco de elección actual.
   - *P*recio, que debería tener en cuenta los costes indirectos y directos (v. sección «Diez preguntas que deben plantearse sobre un análisis económico»).
7. Evalúe rigurosamente la evidencia, prestando especial atención a la potencia (tamaño muestral) y la calidad metodológica de los ensayos clínicos, así como al uso de criterios de valoración indirectos. Aplique la lista de comprobación CONSORT (tabla 6.1). No acepte argumentos teóricos a favor del fármaco (p. ej., semivida más larga) sin una evidencia directa de que esto se refleje en un beneficio clínico.
8. No acepte la novedad de un producto como argumento para usarlo. En realidad, existen buenos argumentos científicos para hacer lo contrario.

9. Rechace probar el producto a través de muestras o participando en «estudios de investigación» a pequeña escala no controlados.
10. Registre por escrito el contenido de la entrevista y repase sus notas si el representante solicita otra cita.

Bibliografía

1 Godlee F. Doctors and the drug industry. BMJ 2008;336 doi: http://dx.doi.org/10.1136/bmj.39444.472708.47.
2 Hollon MF. Direct-to-consumer advertising. JAMA:The Journal of the American Medical Association 2005;**293**(16):2030-3.
3 Liang BA, Mackey T. Direct-to-consumer advertising with interactive internet media global regulation and public health issues. JAMA: The Journal of the American Medical Association 2011;**305**(8):824-5.
4 Kaphingst KA, Dejong W, Rudd RE, et al. *A content analysis of direct-to-consumer television prescription drug advertisements*. *Journal ofHealthCommunication: International Perspectives.* 2004;**9**(6):515-28.
5 Brody H. The company we keep: why physicians should refuse to see pharmaceutical representatives. The Annals of Family Medicine 2005;**3**(1):82-5.
6 O'brien M, Rogers S, Jamtvedt G, et al. Educational outreach visits: effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviews (Online) 2007;**4**(4):1-6.
7 Kravitz RL, Epstein RM, Feldman MD, et al. Influence of patients' requests for direct-to-consumer advertised antidepressants. JAMA:The Journal of the American Medical Association 2005;**293**(16):1995-2002.
8 Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston, USA: Little, Brown and Company; 1985.
9 Goldacre B. *Bad pharma: how drug companies mislead doctors and harm patients*. London, Fourth Estate: Random House Digital Inc; 2013.
10 Rajagopal S. The placebo effect. Psychiatric Bulletin 2006;**30**(5):185-8.
11 Price DD, Finniss DG, Benedetti F. A comprehensive review of the placebo effect: recent advances and current thought. Annual Review of Psychology 2008;**59**:565-90.
12 Gøtzsche PC, Liberati A, Torri V, et al. Beware of surrogate outcome measures. International Journal of Technology Assessment in Health Care 1996;**12**(02):238-46.
13 Connolly SJ. Use and misuse of surrogate outcomes in arrhythmia trials. Circulation 2006;**113**(6):764-6.
14 Guermazi A, Hayashi D, Roemer FW, et al. Osteoarthritis: a review of strengths and weaknesses of different imaging options. Rheumatic Diseases Clinics of North America 2013;**39**(3):567-91.
15 Messerli FH, Staessen JA, Zannad F. Of fads, fashion, surrogate endpoints and dual RAS blockade. European Heart Journal 2010;**31**(18):2205-8.
16 Harel Z, Gilbert C, Wald R, et al. The effect of combination treatment with aliskiren and blockers of the renin–angiotensin system on hyperkalaemia and acute kidney injury: systematic review and meta-analysis. BMJ: British Medical Journal 2012;**344**:e42.
17 Bero L. Industry sponsorship and research outcome: a Cochrane review. JAMA Internal Medicine 2013;**173**(7):580-1.

18 Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Annals of Internal Medicine 2010;**152**(11): 726-32.
19 Turner L, Shamseer L, Altman DG, et al. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Systematic Reviews 2012;**1**:60.
20 Chalmers I, Glasziou P, Godlee F. All trialsmust be registered and the results published. BMJ: British Medical Journal 2013;**346**(7890):f105.
21 Herxheimer A. Getting good value fromdrug reps. Drug andTherapeutics Bulletin 1983;**21**:13-5.
22 Montori VM, Jaeschke R, Schünemann HJ, et al. Users' guide to detecting misleading claims in clinical research reports. BMJ: British Medical Journal 2004;**329**(7474):1093.

Capítulo 7 Artículos que describen ensayos de intervenciones complejas

Intervenciones complejas

En la sección «¿Qué información debería proporcionar un artículo que describa un ensayo controlado aleatorizado?: declaración CONSORT» definí una intervención simple (p. ej., un fármaco) como aquella que está bien delimitada (es decir, es fácil describir en qué consiste la intervención) y que se presta a un diseño de investigación que compara un «grupo con intervención» frente a un «grupo sin intervención». Una intervención compleja es aquella que no está bien delimitada (es decir, es difícil describir con precisión en qué *consiste* la intervención) y que plantea problemas a los investigadores a la hora de implementarla. Las intervenciones complejas suelen constar de múltiples componentes interactivos y pueden actuar a más de un nivel (p. ej., tanto individual como organizacional). A continuación se muestran ejemplos de intervenciones complejas:

- Asesoramiento o educación para pacientes.
- Educación o formación para el personal sanitario.
- Intervenciones que buscan una implicación activa y permanente del participante (p. ej., actividad física, intervenciones dietéticas, grupos de apoyo o terapia psicológica realizada en persona o a través de internet).
- Intervenciones institucionales destinadas a fomentar la práctica basada en la evidencia (p. ej., auditoría y retroalimentación), que se describen con más detalle en el capítulo 15.

El profesor Penny Hawe y cols.[1] han afirmado que una intervención compleja puede contemplarse como un «núcleo teórico» (los componentes que la convierten en lo que es y que, por lo tanto, los investigadores deben aplicar fielmente) y unas características adicionales no centrales que pueden (y deberían) adaptarse con flexibilidad a las necesidades y circunstancias de cada contexto. Por ejemplo, si la intervención consiste en proporcionar retroalimentación a los médicos acerca de la fidelidad con la que siguen las guías basadas en la evidencia sobre la hipertensión, el *núcleo* de la intervención podría ser la información sobre la proporción de pacientes que lograron el nivel de presión arterial recomendado por las guías en un período de tiempo determinado. Los elementos no centrales pueden ser el modo en el que se da la información (de palabra, por carta y por

correo electrónico), si la retroalimentación se proporciona en forma numérica o como un diagrama o un gráfico circular, si se ofrece de forma confidencial o en un contexto de enseñanza de grupo, etcétera.

Las intervenciones complejas suelen requerir una fase de desarrollo para que se puedan optimizar los diferentes componentes antes de evaluarse en un ensayo controlado aleatorizado a gran escala. Por lo general, el diseño de la intervención incluye una fase de *desarrollo* inicial de entrevistas u observaciones cualitativas y, en ocasiones, una pequeña encuesta para averiguar lo que la gente consideraría aceptable, que se aplica al diseño de la intervención. Esto se sigue de un *ensayo piloto* a pequeña escala (que en realidad es una «prueba general» de un ensayo a gran escala, en el que un pequeño número de participantes se asignan al azar para ver qué problemas prácticos y operativos surgen) y, por último, del ensayo definitivo completo[2].

A continuación, se presenta un ejemplo. Una de mis estudiantes de doctorado quería analizar el impacto de las clases de yoga sobre el control de la diabetes para lo que, inicialmente, dedicó un tiempo a entrevistar tanto a personas con diabetes como a profesores de yoga que trabajaban con alumnos que tenían diabetes. Diseñó un pequeño cuestionario para preguntar a la gente con diabetes si estaban interesados en el yoga y observó que sólo algunas personas lo estaban. Todo esto formaba parte de su *fase de desarrollo.* La literatura de investigación previa sobre el uso terapéutico del yoga le dio algunas orientaciones sobre los elementos centrales de la intervención; por ejemplo, parecía haber buenas razones teóricas para centrarse en ejercicios de tipo relajación en lugar de en posturas de fuerza o flexibilidad con mayores exigencias físicas.

Las entrevistas y cuestionarios iniciales de mi alumna le proporcionaron una gran cantidad de información útil, que empleó para diseñar los elementos no centrales de la intervención de yoga. Sabía, por ejemplo, que sus participantes potenciales eran reacios a viajar muy lejos de su casa, que no querían asistir a clase más de dos veces a la semana, que el subgrupo más interesado en probar el yoga correspondía a quienes se habían jubilado recientemente (60-69 años) y que muchos de los posibles participantes se describían a sí mismos como «poco flexibles» y estaban preocupados por no excederse con los estiramientos. Toda esta información le ayudó a diseñar los detalles de la intervención, como lo que cada participante debería hacer, dónde, con qué frecuencia, con quién, durante cuánto tiempo y con qué materiales o instrumentos.

Nuestra decepción llegó cuando probamos la intervención compleja cuidadosamente diseñada en un ensayo controlado aleatorizado y no tuvo ningún impacto sobre el control de la diabetes, comparada con el grupo control constituido por pacientes que estaban en la lista de espera[3]. En la sección de discusión del artículo donde se presentaban los resultados del ensayo sobre el yoga, ofrecimos dos interpretaciones alternativas. Según la primera, al contrario de lo que se había observado en estudios no aleatorizados previos, el yoga no tiene ningún efecto sobre el control de la diabetes. Según la segunda interpretación, el yoga puede tener un impacto, pero a pesar de nuestros esfuerzos en la fase de

desarrollo, la intervención compleja estaba *inadecuadamente optimizada.* Por ejemplo, muchas personas tuvieron dificultades para participar en el grupo y varias personas de cada clase no hicieron los ejercicios porque les resultaban «demasiado difíciles». Además, aunque los profesores de yoga pusieron gran empeño en las clases que se impartían dos veces a la semana y se proporcionó a las personas una cinta de audio y una esterilla de yoga para llevar a casa, no se hizo hincapié en que los participantes debían practicar sus ejercicios todos los días. Posteriormente descubrimos que casi ninguno de ellos hizo los ejercicios en casa.

Por tanto, para *optimizar* el yoga como una intervención compleja en la diabetes, podríamos considerar medidas como: (i) hacer que un médico o enfermera lo «prescribiese» de modo que el paciente estuviese más motivado para asistir a todas las clases, (ii) colaborar con los profesores de yoga para diseñar ejercicios especiales para las personas mayores inseguras que no pueden realizar ejercicios de yoga convencionales y (iii) determinar con mayor precisión los ejercicios que deberían realizarse en casa.

Este ejemplo muestra que cuando un ensayo de una intervención compleja tiene resultados negativos, no demuestra necesariamente que todas las adaptaciones de esta intervención serán ineficaces en todos los contextos. Por el contrario, tiende a estimular a los investigadores a replantear la cuestión y preguntarse cómo se puede perfeccionar la intervención y adaptarla para que sea más probable que funcione. Debe tenerse en cuenta que, dado que nuestra intervención de yoga requiere más trabajo, no pasamos directamente al ensayo controlado aleatorizado a gran escala sino que volvimos a la fase de desarrollo para tratar de perfeccionar la intervención.

Diez preguntas que deben plantearse sobre un artículo que describa una intervención compleja

En 2008, el Medical Research Council elaboró una guía actualizada para evaluar las intervenciones complejas, que se publicaron resumidas en el *British Medical Journal*[2]. Las preguntas que aparecen a continuación sobre cómo evaluar un artículo que describa una intervención compleja, se basan en dicha guía.

Primera pregunta: ¿Cuál es el problema para el que esta intervención compleja se considera una posible solución?

Es demasiado fácil basar un estudio de una intervención compleja en una serie de suposiciones indiscutidas. ¿Los adolescentes beben demasiado alcohol y tienen demasiadas relaciones sexuales sin protección, por lo que seguro que se necesitan programas educativos para exponerles los peligros de esta conducta? Es una conclusión incorrecta. Es posible que el problema sea el consumo de alcohol por parte de los adolescentes o las conductas sexuales de riesgo, pero la causa subyacente de este problema tal vez no sea la ignorancia, sino (por ejemplo) la presión de grupo y los mensajes de los medios de comunicación. Si se tiene en cuenta cuál es el problema preciso,

seremos capaces de analizar críticamente si la intervención se ha diseñado (de forma explícita o inadvertida) basándose en una teoría de acción apropiada (v. la cuarta pregunta).

Segunda pregunta: ¿Qué se hizo en la fase de desarrollo de la investigación para documentar el diseño de la intervención compleja?

No hay reglas fijas sobre lo que debería hacerse en una fase de desarrollo, pero los autores deberían indicar claramente lo que hicieron y justificarlo. Si la fase de desarrollo incluyó investigación cualitativa (como suele suceder), se debe leer el capítulo 12 para obtener una orientación detallada sobre cómo evaluar estos artículos. Si se utilizó un cuestionario, se debe leer el capítulo 14. Cuando se haya evaluado el trabajo empírico utilizando las listas de comprobación apropiadas para el(los) diseño(s) del estudio, se deberá valorar cómo se usaron estos resultados para documentar el diseño de la intervención. Uno de los componentes de la fase de desarrollo será identificar una población diana y quizás dividirla en subpoblaciones (p. ej., por edad, sexo, origen étnico, nivel educativo o estatus de la enfermedad), cada una de las cuales podría requerir una adaptación específica de la intervención.

Tercera pregunta: ¿Cuáles eran los componentes centrales y no centrales de la intervención?

Esta pregunta puede replantearse de otro modo: (i) ¿cuáles son los elementos que deberían estandarizarse para que siguieran siendo iguales con independencia de dónde se llevase a cabo la intervención? y (ii) ¿cuáles son los elementos que deberían adaptarse al contexto y al entorno? Los autores deben indicar claramente qué aspectos de la intervención se deben estandarizar y cuáles deben adaptarse a las contingencias y las prioridades de cada contexto. Una intervención compleja poco estandarizada puede dar lugar a pocos hallazgos generalizables, mientras que una demasiado estandarizada puede ser inviable en ciertos contextos y, por lo tanto, puede subestimar la eficacia potencial de los elementos centrales. La decisión sobre qué es central y qué no lo es debe basarse en los resultados de la fase de desarrollo.

La intervención de control debe analizarse con el mismo detalle que la intervención experimental. Si la intervención de control consiste en «no hacer nada» (o en una lista de espera), debe describirse la actuación que los participantes del grupo control del ensayo no van a recibir en comparación con los del grupo de intervención. Lo más probable es que el grupo de control reciba una serie de medidas que consisten (por ejemplo) en una evaluación inicial, varias visitas de revisión, algunos consejos básicos y tal vez un folleto o un teléfono de ayuda.

Es fundamental definir lo que se ofrece al grupo de control si el ensayo evalúa un nuevo paquete de asistencia controvertido y costoso. En un ensayo reciente sobre telemedicina denominado *Whole Systems Demostrator,* varios analistas interpretaron que los resultados mostraban que los sistemas de telemedicina instalados en los hogares de las personas dan lugar a un uso significativamente

menor de los servicios hospitalarios y a mejores tasas de supervivencia (aunque con un alto coste unitario)[4]. Sin embargo, el grupo de intervención recibió en realidad una combinación de dos intervenciones: el equipo de telemedicina *y* llamadas telefónicas periódicas de una enfermera. El grupo de control no recibió el equipamiento de telemedicina, pero tampoco las llamadas telefónicas de la enfermera. Puede que fuese el contacto humano, y no la tecnología, lo que marcó la diferencia y resulta frustrante no poder saberlo. En mi opinión, el diseño del estudio era defectuoso, pues no nos dice si la telemedicina es eficaz o no.

Cuarta pregunta: ¿Cuál era el mecanismo de acción teórico de la intervención?

Los autores de un estudio sobre una intervención compleja deberían indicar de forma explícita cuál es el supuesto modo de acción de la intervención, incluyendo una descripción de cómo encajan entre sí los distintos componentes. Esta descripción puede cambiar a medida que los resultados de la fase de desarrollo se analizan y se incorporan para perfeccionar la intervención.

No siempre es evidente por qué (o por qué no) funciona una intervención, sobre todo si consta de múltiples componentes dirigidos a diferentes niveles (p. ej., individuo, familia y organización). Hace unos años, revisé las secciones cualitativas de los ensayos de investigación sobre programas de alimentación escolares para niños desfavorecidos[5]. En 19 estudios, en todos los cuales se había evaluado esta intervención compleja en un ensayo controlado aleatorizado (v. revisión Cochrane y el metaanálisis relacionados[6]), encontré un total de seis mecanismos diferentes por los que esta intervención podría haber mejorado el estado nutricional, el rendimiento escolar o ambos: corrección a largo plazo de las deficiencias nutricionales, alivio a corto plazo del hambre, sentimiento de los niños de ser valorados y cuidados, reducción del absentismo, mejora de la dieta en el hogar inspirada en la mejora de la dieta escolar, y mejora del nivel educativo en una generación que aumentaba el poder adquisitivo y, por lo tanto, reducía el riesgo de pobreza en la próxima generación.

Cuando se realiza la evaluación crítica de un artículo sobre una intervención compleja, habrá que determinar si los mecanismos propuestos por los autores son adecuados. El sentido común es un buen punto de partida al respecto, al igual que una conversación entre un grupo de médicos expertos y usuarios de servicios. Si los autores no describen explícitamente el mecanismo de acción, tal vez haya que deducirlo de forma indirecta. En la sección «Evaluación de las revisiones sistemáticas» se describe una revisión de Grol y Grimshaw[7], que demostró que sólo el 27% de los estudios sobre la aplicación de la evidencia incluían una teoría explícita del cambio.

Quinta pregunta: ¿Qué medidas de resultado se utilizaron y eran sensibles?

Cuando se trata de una intervención compleja, una única medida de resultado tal vez no refleje todos los efectos importantes que puede tener la intervención. Mientras que un ensayo de un fármaco frente a un placebo para el tratamiento de la diabetes suele tener una única medida de resultado principal (por lo

general, el análisis de la HbA1c sanguínea) y tal vez varias medidas de resultado secundarias (índice de masa corporal, riesgo cardiovascular global y calidad de vida), un ensayo de una intervención educativa puede tener múltiples resultados, cada uno de los cuales es importante de distintas maneras. Además de los marcadores de control de la diabetes, de riesgo cardiovascular y de calidad de vida, sería importante saber si los profesionales sanitarios consideraron que la intervención educativa era aceptable y factible de administrar, si las personas acudieron a las sesiones, si los conocimientos de los participantes cambiaron, si se modificó su conducta de autocuidado, si la organización pasó a estar más centrada en el paciente, si las llamadas a una línea de ayuda aumentaron o disminuyeron, etcétera.

Una vez respondidas las primeras cinco preguntas, deberíamos ser capaces de realizar un resumen de lo planteado hasta el momento en términos de población, intervención, comparación y resultados, aunque es probable que sea más extenso que el resumen equivalente de una intervención sencilla.

Sexta pregunta: ¿Cuáles fueron los hallazgos?

Esta pregunta es aparentemente simple aunque debe tenerse en cuenta, como se indica en la pregunta cinco, que una intervención compleja puede tener un impacto significativo en algunas de las medidas de resultado, pero carecer de dicho impacto en otras. Este tipo de hallazgos requieren una interpretación cuidadosa. Los ensayos sobre intervenciones de autocuidado (en las que se enseña a personas con enfermedades crónicas a manejar su afección mediante la modificación de su estilo de vida y el ajuste de su medicación en función de los síntomas, o realizando pruebas en su domicilio sobre el estatus de la enfermedad) consideran de forma generalizada que son eficaces. Sin embargo, en realidad estos programas pocas veces modifican la evolución de la enfermedad subyacente ni hacen que las personas vivan más tiempo; simplemente hacen que las personas se sientan más seguras en el manejo de su enfermedad[8,9]. Sentirse mejor acerca de la enfermedad crónica que se padece puede ser un resultado importante por sí mismo, pero hay que ser muy precisos sobre lo que consiguen las intervenciones complejas (y sobre lo que no consiguen) a la hora de evaluar los hallazgos de los ensayos.

Séptima pregunta: ¿Qué evaluación del proceso se realizó y cuáles fueron sus principales conclusiones?

Una evaluación del proceso es (sobre todo) un estudio cualitativo que se realiza en paralelo a un ensayo controlado aleatorizado y que recopila información sobre los problemas prácticos a los que se enfrenta el personal de primera línea que trata de implementar la intervención[10]. En el estudio sobre el yoga en la diabetes, por ejemplo, los investigadores (uno de los cuales era una estudiante de medicina que realizaba su proyecto de licenciatura) acudieron a las clases de yoga, entrevistaron a los pacientes y a los profesores, escribieron las actas de las reuniones de planificación y, en general, plantearon la pregunta: «¿qué tal va?». Un hallazgo crucial de esta evaluación fue comprobar que algunos de los centros de yoga eran inapropiados. Sólo si en realidad se estaba allí cuando

se impartía la clase de yoga podía constatarse que era imposible relajarse y meditar en un centro de ocio público en el que se oían constantemente ruidosos anuncios por megafonía. De manera más general, las evaluaciones de proceso recogerán las opiniones de los participantes y del personal sobre cómo perfeccionar la intervención y/o de por qué tal vez no esté funcionando según lo previsto.

Octava pregunta: Si los resultados fueron negativos, ¿hasta dónde se puede explicar esto por el fracaso a la hora de aplicar la intervención y/o por una optimización inadecuada de la misma?

Esta pregunta se desprende de la evaluación del proceso. En mi revisión de los programas de alimentación escolar (v. cuarta pregunta), muchos estudios tenían resultados negativos y, al leer los diversos artículos, mi equipo pensó en varias explicaciones de por qué la alimentación escolar podría *no* mejorar el crecimiento ni el rendimiento escolar[5]. Por ejemplo, es posible que no se tomasen los alimentos ofrecidos o que proporcionase muy poca cantidad de los nutrientes esenciales; tal vez los alimentos consumidos tenían una baja biodisponibilidad en niños desnutridos (p. ej., no se absorbían debido a edema intestinal); quizás hubiese una reducción compensatoria de la ingesta de alimentos fuera de la escuela (p. ej., la cena se daba a otro miembro de la familia si se sabía que el niño había sido alimentado en la escuela); tal vez la suplementación se había producido demasiado tarde en el desarrollo del niño, o quizá el programa no se aplicó según lo previsto (p. ej., en un estudio, algunos niños del grupo control recibieron suplementos alimenticios porque el personal de primera línea pensaba, probablemente con razón, que no era ético alimentar a la mitad de los niños con hambre en una clase, pero no a la otra mitad).

Novena pregunta: Si los hallazgos variaron entre los diferentes subgrupos, ¿en qué medida lo han explicado los autores perfeccionando su teoría del cambio?

¿La intervención mejoró los resultados en las mujeres pero no en los varones? ¿En las personas con alto nivel educativo de clase media, pero no en las personas sin educación o de la clase trabajadora? ¿En los centros de atención primaria, pero no en la atención secundaria? ¿O en Manchester, pero no en Delhi? En tal caso, hay que preguntarse por qué. Este interrogante de «¿por qué?» es otra valoración personal. Dado que es una cuestión de interpretar los resultados en su contexto, no se puede responder aplicando un algoritmo técnico o una lista de comprobación. Al leer la sección de discusión del artículo, deberíamos encontrar la explicación de los autores de por qué el subgrupo X se benefició pero el subgrupo Y no. También deberían haber perfeccionado su teoría del cambio para tener en cuenta estas diferencias. Por ejemplo, los estudios sobre los programas de alimentación escolar mostraron un beneficio (global) estadísticamente mayor en los niños más pequeños, lo que hizo que los autores de estos estudios sugiriesen que existe una ventana crítica de desarrollo tras la cual incluso los suplementos con un valor nutricional alto tienen un impacto limitado sobre el crecimiento o el rendimiento[5,6]. Para resaltar

otra de mis áreas de interés, me atrevo a afirmar que una de las principales áreas de crecimiento en la investigación secundaria en los próximos años será determinar qué actuaciones son eficaces en cada tipo de pacientes en lo referente a la educación y el apoyo del autocuidado en distintas enfermedades crónicas.

Décima pregunta: ¿Qué otras investigaciones son necesarias según los autores, y están justificadas?

Quien haya leído este capítulo, sabrá que las intervenciones complejas son multifacéticas, ricas en matices y capaces de influir en múltiples resultados. Los autores que presenten estudios de este tipo de intervenciones tienen la responsabilidad de describir cómo ha influido su estudio en el campo general de la investigación. No deberían limitarse a concluir que «se necesitan más investigaciones» (un corolario inevitable de cualquier estudio científico) sino que deberían indicar *dónde* podrían centrarse mejor los esfuerzos de investigación. De hecho, una de las conclusiones más útiles podría ser una descripción de las áreas donde *no* se necesitan más investigaciones. Los autores deberían indicar, por ejemplo, si la próxima etapa debería ser una investigación cualitativa, un ensayo nuevo y más amplio, o incluso un análisis más detallado de los datos ya recopilados.

Bibliografía

1 Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised controlled trial be? BMJ: British Medical Journal 2004;**328**(7455):1561-3.

2 Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ: British Medical Journal 2008;**337**:a1655.

3 Skoro-Kondza L, Tai SS, Gadelrab R, et al. Community based yoga classes for type 2 diabetes: an exploratory randomised controlled trial. BMC Health Services Research 2009;**9**(1):33.

4 Steventon A, Bardsley M, Billings J, et al. Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial. BMJ: British Medical Journal 2012;**344**:e3874 doi: 10.1136/ bmj.e3874.

5 Greenhalgh T, Kristjansson E, Robinson V. Realist review to understand the efficacy of school feeding programmes. BMJ: British Medical Journal 2007;**335**(7625):858-61 doi: 10.1136/bmj.39359.525174.AD.

6 Kristjansson EA, Robinson V, Petticrew M, et al. School feeding for improving the physical and psychosocial health of disadvantaged elementary school children. Cochrane Database of Systematic Reviews (Online) 2007;(1) CD004676 doi: 10.1002/14651858.CD004676.pub2.

7 Grol R, Grimshaw J. Frombest evidence to best practice: effective implementation of change in patients' care. The Lancet 2003;**362**(9391):1225-30.

8 Foster G, Taylor S, Eldridge S, et al. Self-management education programmes by lay leaders for people with chronic conditions. Cochrane Database of Systematic Reviews (Online) 2007;**4**(4):1-78.

9 Nolte S, Osborne RH. A systematic review of outcomes of chronic disease self-management interventions. Quality of life research 2013;**22**:1805-16.
10 Lewin S, Glenton C, Oxman AD. Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. BMJ: British Medical Journal 2009;**339**:b3496.

Capítulo 8 Artículos que describen pruebas diagnósticas o de cribado

Diez hombres en el banquillo

Para los lectores a quienes les resulte nuevo el concepto de validación de las pruebas diagnósticas y a quienes las explicaciones algebraicas («llamaremos a este valor *x*…») les dejen fríos, el siguiente ejemplo puede serles de ayuda. Diez hombres (para los puristas de la igualdad de género, se asumirá que «hombres» significa hombres o mujeres) están a la espera de juicio por asesinato. En realidad, sólo tres de ellos han cometido un asesinato; los otros siete son inocentes de cualquier delito. Un jurado oye cada caso y declara a seis de estos hombres culpables de asesinato. Dos de los condenados son verdaderos asesinos. Cuatro hombres son encarcelados injustamente. Un asesino queda libre.

Esta información se puede expresar en lo que se denomina *tabla de dos por dos* (fig. 8.1). Se debe tener en cuenta que la «verdad» (es decir, si cada hombre *realmente* cometió un asesinato o no) se expresa a lo largo de la fila del título horizontal, mientras que el veredicto del jurado (que puede reflejar la verdad o no) se expresa en la fila de título vertical.

Se debe observar que estas cifras, si son típicas, reflejan una serie de parámetros de este jurado en particular:

(a) Este jurado identifica correctamente a dos de cada tres asesinos verdaderos.
(b) Absuelve correctamente a tres de cada siete personas inocentes.
(c) Si este jurado ha declarado a una persona culpable, sólo hay una posibilidad entre tres de que la persona sea en realidad un asesino.
(d) Si este jurado ha declarado a una persona inocente, tiene una probabilidad de 3/4 de ser realmente inocente.
(e) En 5 casos de cada 10, el jurado obtiene el veredicto correcto.

Estos cinco parámetros constituyen, respectivamente, la sensibilidad, la especificidad, el valor predictivo positivo, el valor predictivo negativo y la precisión de la actuación de este jurado. En este capítulo se analizan estos cinco parámetros aplicados a las pruebas diagnósticas (o de cribado) en comparación con un diagnóstico «verdadero» o patrón oro. En la sección «Cocientes de verosimilitudes» también se introduce un sexto parámetro, un poco más complicado (pero muy útil), de una prueba diagnóstica, el cociente de verosimilitudes *(likelihood ratio)*. (Después de leer el capítulo completo, repase esta sección. Una vez leído el capítulo,

		Estatus delictivo real	
		Asesino	No asesino
Veredicto del jurado	«Culpable»	Condenado correctamente **2 hombres**	Condenado incorrectamente **4 hombres**
	«Inocente»	**1 hombre** Absuelto incorrectamente	**3 hombres** Absuelto correctamente

Figura 8.1 Tabla de 2 × 2 que muestra el resultado de un juicio de 10 hombres acusados de asesinato.

debería poder calcular que el cociente de verosimilitudes de un veredicto positivo del jurado del ejemplo antes mencionado es de 1,17 y el de uno negativo, de 0,78. En caso contrario, no hay que preocuparse: muchos médicos eminentes no tienen ni idea de qué es el cociente de verosimilitudes.)

Validación de las pruebas diagnósticas frente a un patrón oro

Nuestro limpiacristales me comentó una vez que últimamente sentía mucha sed y que había solicitado a su médico de cabecera que le hiciese la prueba de la diabetes, que es una enfermedad frecuente en su familia. La enfermera de la consulta de su médico le había pedido que llevase una muestra de orina e introdujo una tira reactiva en ella. La tira se puso de color verde, lo que significaba, por lo visto, que no había azúcar (glucosa) en la orina. Según la enfermera, eso significaba que no tenía diabetes.

Me costó mucho explicarle que el resultado de la prueba no significaba necesariamente que no tuviera diabetes, al igual que un veredicto de culpabilidad no convierte *obligatoriamente* a alguien en un asesino. La definición de la diabetes, según la Organización Mundial de la Salud (OMS), es una concentración sanguínea de glucosa mayor de 7 mmol/l en ayunas, o mayor de 11,1 mmol/l 2 horas después de una carga de glucosa oral de 100 g (la tan temida «prueba de tolerancia a la glucosa», donde el participante tiene que ingerir hasta la última gota de una desagradable bebida de glucosa y esperar 2 horas para que le realicen un análisis de sangre)[1]. Estos valores deben obtenerse en dos ocasiones distintas si la persona no tiene síntomas, pero sólo en una ocasión si tiene los síntomas típicos de la diabetes (sed, poliuria, etc.).

Estos criterios estrictos pueden considerarse el *patrón oro* para el diagnóstico de la diabetes. Dicho de otro modo, un paciente que cumpla los criterios de la OMS se puede clasificar como diabético y, si no los cumple, no se puede (aunque debe señalarse que las definiciones oficiales de lo que es y no es una enfermedad

cambian con frecuencia y, de hecho, cada vez que escribo una nueva edición de este libro, tengo que ver si las que yo he citado se han modificado a la luz de la nueva evidencia). No puede decirse lo mismo para la inmersión de una tira reactiva en una muestra de orina puntual. Por un lado, una persona puede ser verdaderamente diabética, pero tener un umbral renal alto, es decir, los riñones conservan la glucosa mucho mejor que los de la mayoría de la gente, por lo que su nivel sanguíneo de glucosa tendría que ser mucho más alto para que la glucosa apareciese en la orina. Alternativamente, se puede ser una persona por lo demás sana con un umbral renal *bajo*, de modo que la glucosa pase a la orina incluso cuando no haya ningún exceso en la sangre. De hecho, como cualquier persona con diabetes podría decir, esta enfermedad se asocia con mucha frecuencia con una prueba negativa de glucosa en la orina.

Sin embargo, el uso de una tira reactiva de orina en lugar de la prueba completa de tolerancia a la glucosa tiene muchas ventajas a la hora de «cribar» a las personas para comprobar si tienen diabetes. La prueba es barata, cómoda, fácil de realizar e interpretar, aceptable para los pacientes y proporciona un resultado instantáneo de sí/no. En la vida real, la gente como mi limpiacristales puede no querer someterse a una prueba de tolerancia oral a la glucosa, sobre todo si son trabajadores autónomos y tienen que perder un día de trabajo para hacerlo. Incluso si estuviese dispuesto a someterse a la prueba, el médico de cabecera podría decidir (con o sin razón) que los síntomas del limpiacristales no merecían el coste de esta prueba relativamente sofisticada. Espero que los lectores observen que, aunque la prueba de orina no puede confirmar a ciencia cierta si alguien es diabético, tiene cierta ventaja práctica sobre el patrón oro, y por eso precisamente se utiliza.

Con el fin de evaluar de manera objetiva la utilidad de la prueba de glucosa en orina para la diabetes, habría que seleccionar una muestra de personas (p. ej., 100) y hacer dos pruebas en cada una de ellas: el análisis de orina (prueba de cribado) y la prueba estándar de tolerancia a la glucosa (patrón oro). A continuación, podría comprobarse, en cada persona, si el resultado de la prueba de cribado se corresponde con el patrón oro. Este análisis se denomina *estudio de validación.* Los resultados del estudio de validación podrían expresarse en una tabla de dos por dos (también denominada matriz de dos por dos) como la de la figura 8.2 y sería posible calcular los diversos parámetros de la prueba como en la tabla 8.1, como se hizo con los parámetros del jurado en la sección de «Intervenciones complejas».

Si los valores para los diversos parámetros de una prueba (como la sensibilidad y la especificidad) están dentro de límites razonables, podría afirmarse que la prueba es *válida* (v. la pregunta siete). La validez de los análisis de glucosa en orina para el diagnóstico de la diabetes fue evaluada hace muchos años por Andersson y cols.[2], cuyos datos se han utilizado en el ejemplo de la figura 8.3. El estudio original se realizó con 3.268 participantes, de los cuales 67 se negaron a proporcionar una muestra o no se analizaron adecuadamente por otro motivo. Por razones de simplicidad, se han ignorado estas anomalías y los resultados se expresan en términos de un denominador (número total analizado) de 1.000 participantes.

		Resultado de la prueba patrón oro	
		Positivo para enfermedad ***a + c***	Negativo para enfermedad ***b + d***
Resultado de la prueba de cribado	Prueba positiva ***a + b***	Verdadero positivo ***a***	Falso positivo ***b***
	c + d Prueba negativa	***c*** Falso negativo	***d*** Verdadero negativo

Figura 8.2 Notación de tabla de 2 × 2 para expresar los resultados de un estudio de validación de una prueba diagnóstica o de cribado.

En realidad, estos datos provienen de un estudio epidemiológico realizado para detectar la prevalencia de la diabetes en una población; la validación del análisis de orina era un tema secundario del estudio principal. Si la validación hubiese sido el objetivo principal del estudio, los participantes seleccionados habrían incluido muchas más personas diabéticas, como se mostrará en la pregunta dos[2]. Los lectores que consulten el artículo original también verán que el patrón oro para el diagnóstico de una auténtica diabetes no era la prueba de tolerancia oral a la glucosa sino una serie menos convencional de observaciones. Sin embargo, el ejemplo sirve a su propósito ya que nos proporciona varias cifras para utilizarlas en las ecuaciones que figuran en la última columna de la tabla 8.1. Los parámetros relevantes del análisis de orina para la diabetes se pueden calcular del siguiente modo:

(a) Sensibilidad = $a/(a + c)$ = 6/27 = 22,2%.
(b) Especificidad = $d/(b + d)$ = 966/973 = 99,3%.
(c) Valor predictivo positivo = $a/(a + b)$ = 6/13 = 46,2%.
(d) Valor predictivo negativo = $d/(c + d)$ = 966/987 = 97,9%.
(e) Precisión = $(a + d)/(a + b + c + d)$ = 972/1.000 = 97,2%.
(f) Cociente de verosimilitudes de un resultado positivo de la prueba = sensibilidad/(1 − especificidad) = 22,2/0,7 = 32.
(g) Cociente de verosimilitudes de un resultado negativo de la prueba = (1 − sensibilidad)/especificidad = 77,8/99,3 = 0,78.

A partir de estos parámetros, se puede deducir por qué no es seguro que el limpiacristales no tenga diabetes. Un análisis positivo de glucosa en orina sólo tiene una sensibilidad del 22%, lo que significa que el análisis no detecta casi al 80% de las personas que realmente tienen diabetes. En presencia de los síntomas clásicos y de antecedentes familiares, las posibilidades iniciales del limpiacristales (probabilidad preprueba) de tener la enfermedad son bastante altas y sólo se reducen a alrededor de un 80% (cociente de verosimilitudes negativo, 0,78; v. sección «Cocientes de verosimilitudes») después de un único análisis de orina negativo. A la vista de

Tabla 8.1 Parámetros de una prueba diagnóstica que pueden calcularse mediante la comparación con un patrón oro en un estudio de validación

Parámetro de la prueba	Nombre alternativo	Cuestión analizada por el parámetro	Fórmula (v. fig. 8.1)
Sensibilidad	Tasa de positivos verdaderos (positivos con enfermedad)	¿Cuál es la eficacia de la prueba para detectar a las personas con la enfermedad?	$a/a + c$
Especificidad	Tasa de negativos verdaderos (negativos sanos)	¿Cuál es la eficacia de la prueba para excluir correctamente a las personas sin la enfermedad?	$d/b + d$
Valor predictivo positivo (VPP)	Probabilidad posprueba de una prueba positiva	Si una persona tiene un resultado positivo, ¿cuál es la probabilidad de que tenga la enfermedad?	$a/a + b$
Valor predictivo negativo (VPN)	Indica la probabilidad posprueba de un resultado negativo[a]	Si una persona tiene un resultado negativo, ¿cuál es la probabilidad de que no tenga la enfermedad?	$d/c + d$
Precisión	–	¿Qué proporción de todas las pruebas han dado el resultado correcto (proporción de verdaderos positivos y verdaderos negativos respecto a todos los resultados)?	$a + d/a + b + c + d$
Cociente de verosimilitudes	–	¿Cuánto más probable es que una persona con la enfermedad tenga un resultado positivo que una persona sin ella?	Sensibilidad/(1 − especificidad)

[a]La probabilidad posprueba de un resultado negativo es (1 − VPN).

		Resultado de la prueba patrón oro (prueba de tolerancia a la glucosa)	
		Positivo para diabetes ***27 personas***	Negativo para diabetes ***973 personas***
Resultado del análisis de glucosa en orina	Presencia de glucosa ***13 personas***	Verdadero positivo ***6***	Falso positivo ***7***
	987 personas Ausencia de glucosa	***21*** Falso negativo	***966*** Verdadero negativo

Figura 8.3 Tabla de 2 × 2 que muestra los resultados de un estudio de validación del análisis de glucosa en orina para la diabetes frente al patrón oro de la prueba de tolerancia a la glucosa (basada en Andersson y cols.[2]).

sus síntomas, es evidente que es necesario realizar una prueba más definitiva para la diabetes[3]. Se debe tener en cuenta que, como muestran las definiciones de la tabla 8.1, si el análisis hubiese sido positivo, el limpiacristales tendría buenas razones para estar preocupado porque, aunque la prueba no es muy *sensible* (es decir, no es buena para detectar a las personas con la enfermedad), es bastante *específica* (es buena para descartar a las personas sin la enfermedad).

A pesar de los resultados de estos estudios de hace casi 20 años, el uso de análisis de orina para «descartar la diabetes» sigue siendo sorprendentemente frecuente en algunos lugares. Sin embargo, el argumento académico se ha desplazado en gran medida a la cuestión de si el análisis de sangre de la HbA1c es lo bastante sensible y específico para servir como prueba de cribado de la diabetes[4,5]. Los argumentos se han vuelto mucho más complejos a medida que los epidemiólogos han aportado evidencia sobre la lesión microvascular precoz (subclínica), pero todavía se aplican los principios esenciales de la matriz de 2 × 2 y las cuestiones acerca de los falsos positivos y falsos negativos. En resumen, la prueba es muy eficaz, pero requiere un análisis de sangre y sus costes no son insignificantes.

Los estudiantes a menudo se confunden sobre la dimensión de sensibilidad/especificidad de una prueba y la dimensión de valor predictivo positivo/negativo. Como regla general, la sensibilidad o la especificidad informa acerca de la *prueba en general*, mientras que el valor predictivo informa sobre *lo que significa un resultado particular de la prueba para un paciente individual.* Por lo tanto, la sensibilidad y la especificidad suelen usarse más por parte de epidemiólogos y especialistas en salud pública cuyo trabajo cotidiano consiste en la toma de decisiones acerca de *poblaciones.*

Una mamografía (radiografía de la mama) de cribado podría tener una sensibilidad del 80% y una especificidad del 90% para la detección del cáncer de mama, lo que significa que la prueba detectará el 80% de los cánceres y descartará al 90% de las mujeres sin cáncer. Sin embargo, ahora imaginemos que trabajamos en atención

primaria y una paciente nos consulta para saber el resultado de su mamografía. La pregunta que ella nos planteará (si el resultado de la prueba ha sido positivo) es: «¿cuál es la probabilidad de que tenga cáncer?» o (si ha sido negativo): «¿cuál es la probabilidad de que pueda descartar la posibilidad de tener cáncer?». Muchos pacientes (y demasiados profesionales sanitarios) asumen que el valor predictivo negativo de una prueba es del 100% (es decir, si la prueba es «normal», piensan que no hay probabilidades de que haya enfermedad), pero basta con leer algunos artículos de revistas femeninas («me dijeron que tenía cáncer, pero las pruebas demostraron después que los médicos se habían equivocado») para encontrar ejemplos de mujeres que han asumido que el valor predictivo positivo de una prueba es del 100%.

Diez preguntas que deben plantearse sobre un artículo que pretende validar una prueba diagnóstica o de cribado

A la hora de preparar estos consejos, me he basado en tres fuentes publicadas principales: *Users' Guides to the Medical Literature*[6,7], un artículo más reciente de algunos de los mismos autores[8] y las guías clínicas simples y pragmáticas de Mant[9] para evaluar una prueba. Al igual que muchas de las listas de comprobación que aparecen en este libro, no son más que reglas prácticas generales para el evaluador crítico principiante: se puede consultar una serie mucho más amplia y rigurosa de criterios (que ocupa la abrumadora cifra de 234 páginas), denominada lista de comprobación *QADAS* (*Quality in Diagnostic and Screening tests,* o calidad en pruebas diagnósticas y de cribado) en una revisión reciente realizada por el Health Technology Assessment Programme de Reino Unido[8]. Lucas y cols.[10] han elaborado una lista de comprobación similar, pero no idéntica, a las preguntas que se presentan aquí.

Pregunta uno: ¿Es esta prueba potencialmente relevante para mi práctica?

Ésta es la pregunta del «¿y para qué sirve?», que los epidemiólogos denominan la *utilidad* de la prueba. Aunque esta prueba fuese válida, precisa y fiable al 100%, ¿nos resultaría de ayuda? ¿Identificaría un trastorno tratable? En caso afirmativo, ¿debería utilizarse preferentemente respecto a la prueba que se usa en la actualidad? ¿Podría el médico (o el paciente o el contribuyente) pagarla? ¿Otorgarían los pacientes su consentimiento para realizarla? ¿Cambiaría las probabilidades de los diagnósticos diferenciales lo suficiente para modificar el plan terapéutico? Si las respuestas a estas preguntas son negativas, se puede descartar el artículo sin necesidad de leer más allá del resumen o la introducción.

Pregunta dos: ¿Se ha comparado la prueba con un verdadero patrón oro?

En primer lugar, hay que preguntarse si la prueba se ha comparado con alguna otra prueba. En ocasiones se escriben artículos (que, antiguamente, llegaban incluso a publicarse) en los cuales no se había hecho nada, salvo realizar la nueva prueba en unas docenas de participantes. Esta intervención puede

proporcionar un rango de posibles resultados de la prueba, pero no confirma que los resultados «altos» indiquen la presencia del trastorno en estudio (la enfermedad o el estado de riesgo que nos interesa) o que los resultados bajos indiquen su ausencia.

A continuación, se debe comprobar que la prueba usada como «patrón oro» en el estudio merezca tal denominación. Una buena forma de evaluar un patrón oro es usar las preguntas del ¿y para qué sirve? enumeradas anteriormente. Para muchas enfermedades, no existe una prueba diagnóstica que sea un patrón oro absoluto que nos indique con certeza su presencia o no. No resulta sorprendente que éstas tiendan a ser las enfermedades para las cuales la búsqueda de nuevas pruebas sea más intensa. Por lo tanto, es posible que los autores de estos artículos tengan que elaborar y justificar una combinación de criterios frente a los cuales se debe evaluar la nueva prueba. Un punto específico que se debe comprobar es que la prueba que se está validando (o una variante de la misma) no se utilice para contribuir a la definición del patrón oro.

Pregunta tres: ¿Este estudio de validación incluye un espectro adecuado de participantes?

Si se ha validado un nuevo análisis del colesterol en 100 varones estudiantes de medicina sanos, no podría afirmarse cuál sería la eficacia de la prueba en mujeres, niños, personas mayores, pacientes con enfermedades que eleven intensamente la cifra de colesterol, o incluso en quienes nunca hayan estado en una facultad de medicina. Aunque pocas personas serían tan ingenuas para seleccionar una muestra tan sesgada para su estudio de validación, es sorprendentemente habitual que los estudios publicados no definan el espectro de participantes evaluados en términos de edad, sexo, síntomas y/o gravedad de la enfermedad, así como los criterios de elegibilidad específicos.

Es fundamental definir tanto la gama de participantes como el espectro de la enfermedad que se va a incluir si se quiere que merezca la pena mencionar los valores de los diferentes parámetros de la prueba (es decir, para que puedan extrapolarse a otros contextos). Una prueba diagnóstica en particular puede ser más sensible en las participantes femeninas que en los participantes masculinos, o más sensible en los más jóvenes que en los de mayor edad. Por las mismas razones, los participantes en quienes se verifica cualquier prueba deben incluir a personas que tengan formas tanto leves como graves de la enfermedad, tratadas y no tratadas, así como aquellas que tengan enfermedades diferentes pero que suelen confundirse.

Mientras que la sensibilidad y especificidad de una prueba son prácticamente constantes con independencia de la prevalencia de la enfermedad, los valores predictivos positivos y negativos son muy dependientes de la prevalencia. Esto explica por qué los médicos de atención primaria son (a menudo con razón) escépticos sobre la utilidad de las pruebas desarrolladas exclusivamente en una población de atención secundaria, donde la gravedad de la enfermedad tiende a ser mayor (v. sección ¿Qué pacientes incluye el estudio?), y por qué una buena prueba *diagnóstica* (utilizada por lo general cuando el paciente

tiene algunos síntomas sugestivos de la enfermedad en cuestión) no es necesariamente una buena prueba de *cribado* (utilizada habitualmente en personas sin síntomas, tomadas de una población con una prevalencia mucho menor de la enfermedad).

Pregunta cuatro: ¿Se ha evitado el sesgo de confirmación (verificación)?

Esto es fácil de comprobar. Simplemente significa si «en todas las personas en quienes se realizó la nueva prueba diagnóstica también se utilizó el patrón oro y viceversa». El lector no debería tener ningún problema para detectar el posible sesgo en los estudios en los que el patrón oro sólo se realice en personas que ya han tenido un resultado positivo en la prueba que se está validando. Además, existen varios aspectos más sutiles del sesgo de confirmación o verificación que escapan al alcance de este libro, pero que se describen en los libros de estadística especializados[11].

Pregunta cinco: ¿Se ha evitado el sesgo de expectativa?

El sesgo de expectativa se produce cuando los patólogos y otros profesionales que interpretan las muestras para el diagnóstico están influenciados inconscientemente porque conocen las características particulares del caso; por ejemplo, la presencia de dolor torácico al interpretar un electrocardiograma (ECG). En el contexto de la validación de pruebas diagnósticas frente al patrón oro, esta pregunta significa si «las personas que interpretaron una de las pruebas sabían cuál había sido el resultado de la otra prueba en cada participante individual». Como se ha explicado en la sección «¿Se realizó la evaluación de forma ciega?», todas las evaluaciones deberían realizarse de forma «ciega», es decir, la persona que interpreta la prueba no debería tener ninguna pista sobre cuál es el resultado esperado en ningún caso concreto.

Pregunta seis: ¿Se demostró que la prueba es reproducible en un mismo observador y entre distintos observadores?

Si el mismo observador realiza la misma prueba en dos ocasiones en un participante cuyas características no han cambiado, obtendrá resultados diferentes en una proporción de casos. Todas las pruebas presentan esta característica en cierta medida, pero una prueba con una reproducibilidad del 99% está claramente en una categoría distinta que una con una reproducibilidad del 50%. Entre los factores que pueden contribuir a la escasa reproducibilidad de una prueba diagnóstica se encuentran la precisión técnica de los aparatos, la variabilidad del observador (p. ej., al comparar un color con una tabla de referencia), errores aritméticos, etcétera.

Se debe volver a la sección «¿Se realizó la evaluación de forma ciega?» para recordar el problema de la concordancia interobservador. A la hora de interpretar un mismo resultado, dos personas coincidirán sólo en una proporción de los casos, lo que suele expresarse como la puntuación kappa. Si la prueba en cuestión proporciona los resultados en términos numéricos (como la concentración sérica de colesterol en milimoles por litro), la concordancia interobservador no suele ser un problema. Sin embargo, si la prueba consiste en la lectura de radiografías (como el ejemplo de la mamografía de la sección «¿Se

realizó la evaluación de forma ciega?») o en preguntar a una persona sobre sus hábitos de ingesta de alcohol[10], es importante confirmar que la reproducibilidad interobservador tenga un nivel aceptable.

Pregunta siete: ¿Cuáles son los parámetros de la prueba derivados de este estudio de validación?

Aunque se hubiesen cumplido todas estas normas, la prueba aún podría carecer de utilidad porque ella misma no fuese válida (es decir, su sensibilidad, especificidad y otros parámetros cruciales son demasiado bajos. Esto es justo lo que sucede cuando se utiliza la glucosa en orina para el cribado de la diabetes; v. la sección «Diez preguntas que deben plantearse sobre un artículo que describa una intervención compleja»). Después de todo, si una prueba tiene una tasa de falsos negativos cercana al 80%, es más probable que engañe al médico que ayude al diagnóstico si la enfermedad para la que está diseñada está realmente presente.

No hay verdades absolutas sobre la validez de una prueba de cribado, porque lo que se considera aceptable depende de la enfermedad para la que se está realizando el cribado. Pocos profesionales criticarían una prueba para el daltonismo que tuviese una sensibilidad del 95% y una especificidad del 80%, pero nunca nadie ha muerto de daltonismo. La prueba del talón para la detección del hipotiroidismo congénito, que se realiza en todos los bebés en países desarrollados poco después del nacimiento, tiene una sensibilidad superior al 99%, pero su valor predictivo positivo es sólo del 6% (es decir, detecta casi todos los bebés con la enfermedad a costa de una tasa elevada de falsos positivos)[11], pero esto está justificado. Es mucho más importante detectar a todos y cada uno de los bebés con esta enfermedad tratable que de lo contrario desarrollarían una discapacidad mental profunda que evitar a cientos de padres el estrés relativamente leve de tener que repetir un análisis de sangre a su hijo.

Pregunta ocho: ¿Se indicaron los intervalos de confianza para la sensibilidad, especificidad y otros parámetros de la prueba?

Como se ha explicado en la sección «Probabilidad y confianza», un intervalo de confianza, que se puede calcular para casi todos los aspectos numéricos de un conjunto de resultados, expresa la posible gama de resultados en la que puede encontrarse el valor verdadero. Volviendo al ejemplo del jurado de la sección «Intervenciones complejas», si se hubiese declarado inocente a un único asesino más, la sensibilidad de su veredicto habría bajado del 67% al 33%, y el valor predictivo positivo de la sentencia, del 33% al 20%. Esta gran (y bastante inaceptable) influencia de una sola decisión se debe a que la acción del jurado se validó sólo con 10 casos (¡los intervalos de confianza para los parámetros de este jurado son tan amplios que mi programa informático no permite calcularlos!). Debe recordarse que cuanto mayor sea el tamaño muestral, más estrecho será el intervalo de confianza, por lo que es fundamental buscar los intervalos de confianza si el artículo que se está leyendo describe un estudio sobre una muestra relativamente pequeña. Los lectores que quieran conocer la fórmula para calcular los intervalos de confianza para los parámetros de las

pruebas diagnósticas pueden consultar el excelente *Statistics with Confidence*[12].

Pregunta nueve: ¿Se ha obtenido un «rango de la normalidad» razonable a partir de estos resultados?

Si la prueba proporciona resultados no dicotómicos (continuos), es decir, si proporciona un valor numérico en lugar de un resultado de tipo sí/no, alguien tendrá que decir a partir de qué valor el resultado de la prueba se considerará anormal. Muchas personas han pasado por una experiencia similar cuando se les mide la presión arterial. Siempre se quiere saber si el resultado es adecuado o no, pero a veces el resultado son cifras como 142/92. Si se escogiera el valor 140/90 como punto de corte de una presión arterial alta, ese resultado estaría en la categoría «anormal» a pesar de que el riesgo de presentar problemas debido a la presión arterial difiere muy poco del de una persona con una presión arterial de 138/88. Con un criterio muy sensato, muchos médicos y enfermeras explican a sus pacientes que su presión arterial no está del todo bien, pero no se encuentra en la zona de peligro y les aconsejan volver dentro de tres meses para realizar una nueva determinación. Sin embargo, el médico debe tomar en algún momento la decisión de qué presión arterial *específica* necesita tratamiento con pastillas y cuál no. Cuándo y con qué frecuencia se debe repetir una prueba cuyos resultados están en la zona límite suele detallarse en las guías clínicas; por ejemplo, se pueden consultar los consejos detallados y las controversias actuales sobre el modo de medir la presión arterial[13].

La definición de las zonas de riesgo relativo y absoluto de una variable continua fisiológica o patológica es una ciencia compleja, que debería tener en cuenta la probabilidad real de los resultados adversos que el tratamiento propuesto intenta prevenir. La objetividad de este proceso aumenta considerablemente cuando se usan los cocientes de verosimilitudes (v. sección «Cocientes de verosimilitudes»). En la página 59 de la obra de Sackett y cols.[14] se presenta una descripción amena sobre los diferentes significados posibles de la palabra «normal» en las pruebas diagnósticas.

Pregunta diez: ¿Se ha puesto esta prueba en el contexto de otras pruebas posibles en la secuencia de diagnóstico de la enfermedad?

En general, la presión arterial alta se trata basándose únicamente en la medición de la presión arterial (aunque, como se ha mencionado, las guías recomiendan basar el tratamiento en una serie de mediciones en lugar de en un único valor). Esto se puede comparar con la secuencia que se utiliza para diagnosticar la estenosis de las arterias coronarias. En primer lugar, se selecciona a los pacientes con una historia típica de angina de esfuerzo (dolor torácico con el ejercicio). A continuación, suele realizarse un ECG en reposo, un ECG de esfuerzo y, en algunos casos, una gammagrafía cardíaca para buscar áreas a las que llegue poco oxígeno. En la mayoría de los pacientes, sólo se realiza una angiografía coronaria (la prueba definitiva para detectar la estenosis de las arterias coronarias) después de que se haya obtenido un resultado anormal en estas pruebas preliminares.

Si se escogiera a 100 personas en la calle y se les realizase directamente una angiografía coronaria, la prueba podría mostrar valores predictivos positivos y negativos muy diferentes (e incluso una sensibilidad y especificidad distintas) de los que se obtuvieron en la población más enferma en la que se validó inicialmente. Esto significa que los diversos aspectos de la validez de la angiografía coronaria como prueba diagnóstica carecen prácticamente de sentido a menos que estas cifras se expresen en términos de su aportación al estudio diagnóstico global.

Cocientes de verosimilitudes

En la pregunta nueve se describe el problema de definir un rango normal para una variable continua. En tales circunstancias, puede ser preferible expresar el resultado de la prueba no como «normal» o «anormal», sino en términos de las probabilidades reales de que un paciente tenga el trastorno específico si el resultado de la prueba alcanza un nivel particular. Consideremos, por ejemplo, el uso del análisis del antígeno prostático específico (PSA) para detectar el cáncer de próstata. La mayoría de los varones tendrán cierta cantidad de PSA detectable en la sangre (p. ej., 0,5 ng/ml) y la mayoría de las personas con cáncer de próstata avanzado tendrán niveles muy altos de PSA (mayores de alrededor de 20 ng/ml). Sin embargo, un nivel de PSA de 7,4 ng/ml, por ejemplo, puede corresponder a un varón perfectamente sano o a una persona con cáncer en un estadio precoz. En este caso, no hay un punto de corte nítido entre normal y anormal[15].

Sin embargo, se pueden utilizar los resultados de un estudio de validación del análisis de PSA frente al patrón oro para el cáncer de próstata (la biopsia) para elaborar toda una serie de tablas de dos por dos. Cada tabla usaría una definición diferente de un resultado anormal de PSA para clasificar a los pacientes como «normales» o «anormales». A partir de estas tablas, se podrían generar diferentes cocientes de verosimilitudes asociados con un nivel de PSA por encima de cada punto de corte diferente. Después, si se obtiene un resultado de PSA en la «zona gris», al menos se podría afirmar: «esta prueba no ha demostrado que el paciente tenga cáncer de próstata, pero ha incrementado [o disminuido] las posibilidades de dicho diagnóstico un factor *x*». De hecho, como he mencionado antes, el análisis de PSA no es especialmente eficaz a la hora de distinguir entre la presencia y la ausencia de cáncer, con independencia del valor de corte que se utilice. Dicho de otro modo, no existe un valor de PSA que proporcione un cociente de verosimilitudes especialmente alto para la detección del cáncer. El último consejo consiste en compartir estas incertidumbres con el paciente y dejarle que decida si se somete a la prueba[16].

Aunque el cociente de verosimilitudes es uno de los elementos más complicados de calcular de una prueba diagnóstica, tiene un enorme valor práctico y se está convirtiendo en la forma preferida de expresar y comparar la utilidad de diferentes pruebas. El cociente de verosimilitudes es una prueba especialmente útil para descartar o confirmar un diagnóstico en particular. Por ejemplo, en

Reino Unido, si una persona entra en la consulta del médico sin ningún síntoma en absoluto, se sabe (basándose en algunos estudios epidemiológicos más antiguos) que presenta un 5% de probabilidades de tener anemia ferropénica ya que alrededor de una de cada 20 personas en la población británica tiene esta enfermedad. En el lenguaje de las pruebas diagnósticas, esto significa que la probabilidad preprueba de tener anemia, que es equivalente a la prevalencia de la enfermedad, es de 0,05.

A continuación, si se realiza una prueba diagnóstica para la anemia, como la determinación del nivel sérico de ferritina, el resultado suele hacer que el diagnóstico de la anemia sea más o menos probable. Un nivel sérico de ferritina moderadamente reducido (entre 18 y 45 μg/l) tiene un cociente de verosimilitudes de 3, por lo que las probabilidades de que un paciente con este resultado tenga anemia ferropénica suelen calcularse como 0,05 × 3, o 0,15 (15%). Este valor se denomina *probabilidad posprueba del análisis de ferritina sérica*. (En sentido estricto, los cocientes de verosimilitudes se deberían utilizar con posibilidades y no con probabilidades, pero el método más simple que se muestra aquí ofrece una buena aproximación cuando la probabilidad preprueba es baja. En este ejemplo, una probabilidad preprueba del 5% es igual a unas posibilidades preprueba de 0,05/0,95 o 0,053; una prueba positiva con un cociente de verosimilitudes de 3 se asocia a unas posibilidades posprueba de 0,158, lo que equivale a una probabilidad posprueba del 14%)[17].

En la figura 8.4 se muestra un nomograma, adaptado por Sackett y cols. a partir de un artículo original de Fagan[18], para obtener las probabilidades posprueba cuando se conocen la probabilidad preprueba (prevalencia) y el cociente de verosimilitudes de la prueba. Las líneas A, B y C, trazadas a partir de una probabilidad preprueba del 25% (la prevalencia del tabaquismo en los adultos británicos) son, respectivamente, las trayectorias que pasan por los cocientes de verosimilitudes de 15, 100 y 0,015, correspondientes a tres pruebas diferentes (y algo anticuadas) para detectar si alguien es fumador[19]. En realidad, la prueba de C detecta si la persona es un *no fumador,* pues un resultado positivo en esta prueba conlleva una probabilidad posprueba de sólo el 0,5%.

En resumen, como mencioné al principio de este capítulo, se pueden utilizar ampliamente las pruebas diagnósticas sin tener que recurrir a los cocientes de verosimilitudes. Yo misma los evité durante años. Sin embargo, si se dedica una tarde a familiarizarse con este aspecto de la epidemiología clínica, ese tiempo habrá sido bien empleado.

Reglas de predicción clínica

En la sección anterior se expuso un ejemplo algo farragoso de la prueba del PSA, llegando a la conclusión de que no existe un valor único y nítido que distinga de forma fiable lo «normal» de lo «anormal». Por ello, la estrategia recomendada para evaluar el riesgo de un varón de tener cáncer de próstata es una combinación de varias pruebas, que incluye la evaluación clínica general y un tacto rectal[16].

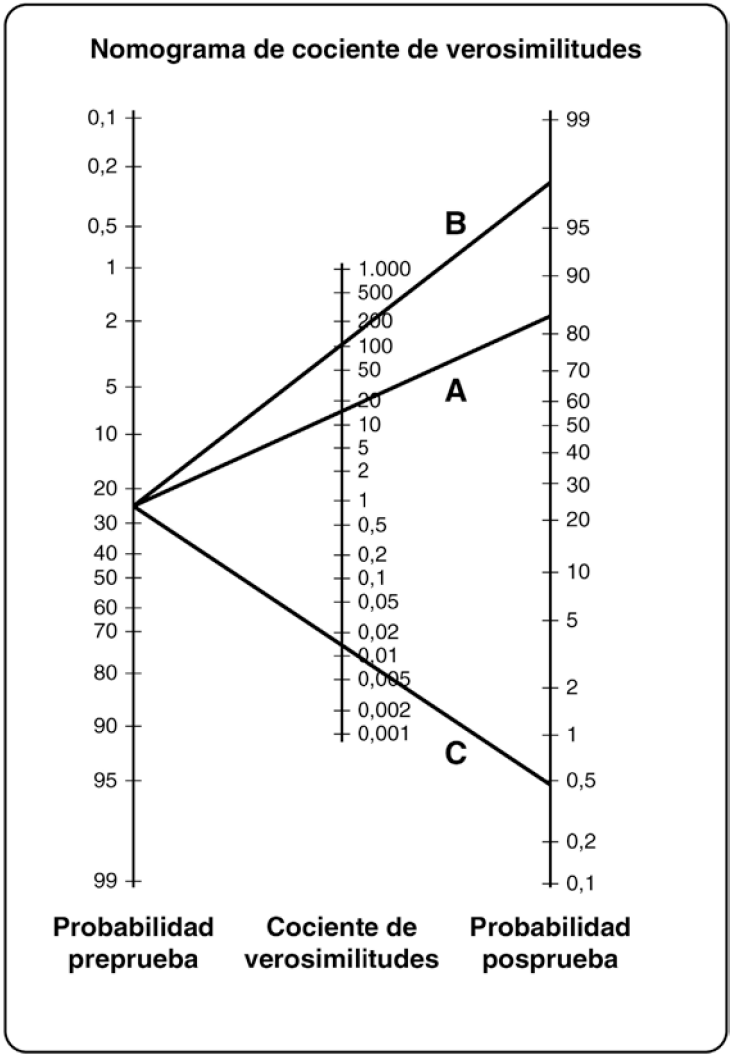

Figura 8.4 Uso de los cocientes de verosimilitudes para calcular la probabilidad posprueba de que alguien sea fumador.

De forma más general, se puede observar por qué los médicos habitualmente tienden a utilizar una combinación de varias pruebas diagnósticas diferentes (que incluyen la exploración física, análisis de sangre, radiografías, etc.) para elaborar una imagen del problema que presenta el paciente. Aunque cualquier prueba individual tiene una frontera difusa entre lo normal y lo anormal, la combinación

de varias pruebas puede afinar la precisión diagnóstica. De este modo, por ejemplo, en una mujer que presente una tumoración en la mama se suelen realizar tres pruebas diferentes, ninguna de las cuales es especialmente útil cuando se utiliza de manera aislada: aspiración con aguja fina, radiografía (mamografía) y ecografía[20]. Más recientemente, los académicos han iniciado un debate acerca de si la lectura informatizada de la mamografía aumenta aún más la precisión de esta triple combinación[21].

Este principio general de hacer varias pruebas y combinarlas se lleva empleando desde hace mucho tiempo en la práctica clínica y recientemente Falk y Fahey lo han actualizado de una forma más estructurada[22]. Mediante el seguimiento de grandes cohortes de pacientes con síntomas específicos y el registro exhaustivo de los resultados de las exploraciones físicas y las pruebas diagnósticas en todos ellos se pueden obtener estimaciones numéricas de la probabilidad de que una persona tenga (o vaya a desarrollar) la enfermedad X en presencia del síntoma A, del signo físico B, de la prueba diagnóstica C, etcétera, o de cualquier combinación de éstos. En los últimos años ha aumentado rápidamente el interés (y las investigaciones) sobre las reglas de predicción clínica, debido en parte a que los avances de la tecnología de la información permiten que los médicos de diferentes centros introduzcan un gran número de pacientes en las bases de datos en línea.

Como Falk y Fahey señalan, hay tres etapas en el desarrollo de una regla de predicción clínica. En primer lugar, la regla se debe desarrollar estableciendo el efecto independiente y combinado sobre el diagnóstico de variables explicativas, como síntomas, signos o pruebas diagnósticas. En segundo lugar, estas variables explicativas deben evaluarse en diferentes poblaciones. Y en tercer lugar, debe efectuarse un análisis de impacto (lo ideal sería realizar un ensayo aleatorizado que midiese el impacto de la aplicación de la regla en un contexto clínico, en términos de evolución del paciente, conducta clínica, uso de recursos, etc.).

En la bibliografía se recogen ejemplos de cómo las reglas de predicción clínica pueden ayudarnos a superar algunas de las dificultades diagnósticas más complejas en la asistencia sanitaria, como la manera de predecir si en un niño con traumatismo craneoencefálico se debe solicitar una tomografía computarizada[23], si un paciente con artritis precoz está desarrollando artritis reumatoide[24], si alguien que toma anticoagulantes presenta un riesgo suficientemente bajo de accidente cerebrovascular como para suspenderlos[25] y qué combinaciones de pruebas predicen mejor si un niño gravemente enfermo presenta algún tipo de problema crítico[26].

Bibliografía

1 World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. Geneva: World Health Organization; 2006 1-50.

2 Andersson D, Lundblad E, Svärdsudd K. A model for early diagnosis of type 2 diabetes mellitus in primary health care. Diabetic Medicine 1993;**10**(2):167-73.

3 Friderichsen B, Maunsbach M. Glycosuric tests should not be employed in population screenings for NIDDM. Journal of Public Health 1997;**19**(1):55-60.
4 Bennett C, Guo M, Dharmage S. HbA1c as a screening tool for detection of type 2 diabetes: a systematic review. Diabetic Medicine 2007;**24**(4):333-43.
5 Lu ZX, Walker KZ, O'Dea K, et al. A1C for screening and diagnosis of type 2 diabetes in routine clinical practice. Diabetes Care 2010;**33**(4):817-9.
6 Jaeschke R, Guyatt G, Sackett DL, et al. Users' guides to the medical literature: III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA: The Journal of the American Medical Association – US Edition 1994;**271**(5):389-91.
7 Guyatt G, Bass E, Brill-Edwards P, et al. Users' guides to the medical literature: III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? JAMA:The Journal of American Medical Association 1994;**271**(9):703-7.
8 Guyatt G, Sackett D, Haynes B. Evaluating diagnostic tests. *Clinical epidemiology: how to do clinical practice research* 2006;**424**:273-322.
9 Mant D. Testing a test: three critical steps. Oxford General Practice Series 1995;**28**:183.
10 Lucas NP, Macaskill P, Irwig L, et al. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). Journal of Clinical Epidemiology 2010;**63**(8):854-61.
11 Lu Y, DendukuriN, Schiller I, et al. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. Statistics in Medicine 2010;**29**(24):2532-43.
12 Altman DG, Machin D, Bryant TN, et al. *Statisticswith confidence: confidence intervals and statistical guidelines*. Bristol: BMJ Books; 2000.
13 Appel LJ. Miller ER,Charleston J. Improving themeasurement of blood pressure: is it time for regulated standards? Annals of InternalMedicine 2011;**154**(12):838-9.
14 Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little Brown and Company; 1985.
15 Holmström B, Johansson M, Bergh A, et al. Prostate specific antigen for early detection of prostate cancer: longitudinal study. BMJ: British Medical Journal 2009;**339**:b3537.
16 Barry M, Denberg T, Owens D, et al. Screening for prostate cancer: a guidance statement from the Clinical Guidelines Committee of the American College of Physicians. Annals of Internal Medicine 2013;**158**:761-9.
17 Guyatt GH, Patterson C, Ali M, et al. Diagnosis of iron-deficiency anemia in the elderly. The American Journal of Medicine 1990;**88**(3):205-9.
18 Fagan TJ. Letter: nomogram for Bayes theorem. The New England Journal of Medicine 1975;**293**(5):257.
19 Moore A, McQuay H, Muir Gray J. How good is that test–using the result. Bandolier. Oxford 1996;**3**(6):6-8.
20 Houssami N, Irwig L. Likelihood ratios for clinical examination, mammography, ultrasound and fine needle biopsy in women with breast problems. The Breast 1998;**7**(2):85-9.
21 Giger ML. Update on the potential of computer-aided diagnosis for breast cancer. Future Oncology 2010;**6**(1):1-4.
22 Falk G, Fahey T. Clinical prediction rules. BMJ: British Medical Journal 2009;**339**:b2899.
23 Maguire JL, Boutis K, Uleryk EM, et al. Should a head-injured child receive a head CT scan? A systematic review of clinical prediction rules. Pediatrics 2009;**124**(1):e145-54.
24 Kuriya B, Cheng CK, Chen HM, et al. Validation of a prediction rule for development of rheumatoid arthritis in patients with early undifferentiated arthritis. Annals of the Rheumatic Diseases 2009;**68**(9):1482-5.

25 Rodger MA, Kahn SR, Wells PS, et al. Identifying unprovoked thromboembolism patients at low risk for recurrence who can discontinue anticoagulant therapy. Canadian Medical Association Journal 2008;**179**(5):417-26.
26 Verbakel JY, Van den Bruel A, Thompson M, et al. How well do clinical prediction rules performin identifying serious infections in acutely ill children across an international network of ambulatory care datasets? BMC Medicine 2013;**11**(1):10.

Capítulo 9 Artículos que resumen otros artículos (revisiones sistemáticas y metaanálisis)

¿Cuando es sistemática una revisión?

Todos recordamos los trabajos que teníamos que escribir en los primeros cursos de la universidad. Deambulábamos por la biblioteca, consultando los índices de libros y revistas. Cuando encontrábamos un párrafo que parecía relevante, lo copiábamos y, si había algo que no encajaba con la teoría que estábamos proponiendo, lo descartábamos. Esto, más o menos, constituye una revisión *periodística:* una revisión de estudios primarios que no se han identificado ni analizado de forma sistemática (es decir, estandarizada y objetiva). Los periodistas cobran en función de la cantidad de texto que escriben en lugar de por cuánto leen o por el grado de análisis crítico al que someten lo que han leído. Esto explica por qué la mayoría de los «nuevos avances científicos» que aparecen en la prensa probablemente queden desacreditados antes de un mes. Una variante frecuente de la revisión periodística es la revisión por invitación, que se elabora cuando un editor le pide a uno de sus amigos que escriba un artículo y que se resume con un título llamativo como: «Revisión por invitación. O, mi experiencia, desde mi punto de vista, escrito por mí usando sólo mis datos y mis ideas, y citando sólo mis publicaciones»[1].

En cambio, una *revisión sistemática* es una revisión de estudios primarios que:

- Contiene una declaración de objetivos, fuentes y métodos.
- Se ha llevado a cabo de una manera explícita, transparente y reproducible (v. fig. 9.1).

Las revisiones sistemáticas más duraderas y fiables, en especial las realizadas por la Colaboración Cochrane (v. sección «Recursos especializados») se actualizan periódicamente para incorporar nueva evidencia.

Como mi colega Paul Knipschild observó hace algunos años, el ganador del premio Nobel Pauling[2] publicó una vez una revisión, basada en referencias seleccionadas de los estudios que apoyaban su hipótesis, para demostrar que la vitamina C curaba el resfriado común. Un análisis más objetivo demostró que, aunque el 50% de ellas sugerían que sí se producía un efecto, una verdadera estimación basada en *todos* los estudios disponibles sugirió que la vitamina C no tenía ningún efecto en absoluto sobre la evolución del resfriado común. Pauling probablemente no tenía la intención deliberada de engañar a sus lectores, pero dado que su entusiasmo por la hipótesis que planteaba superó su objetividad

Figura 9.1 Método para una revisión sistemática.

científica, no advirtió el *sesgo de selección* que influyó a la hora de escoger los artículos. La evidencia muestra que cualquiera que tratase de hacer lo que Pauling llevó a cabo, es decir, la búsqueda en la literatura médica de evidencia que respalde nuestra teoría favorita, haría un trabajo igual de personalista y poco científico[3]. Algunas de las ventajas de la revisión sistemática se resumen en el cuadro 9.1.

En una ocasión, se demostró que los expertos que han estado inmersos en un tema durante años y saben cuál «debería» ser la respuesta son significativamente menos capaces de elaborar una revisión objetiva de la literatura sobre su materia que los no expertos[4]. Esto tendría poca importancia si se pudiera confiar en que la opinión de los expertos fuese congruente con los resultados de las revisiones sistemáticas independientes, pero lo cierto es que en ese momento, en la mayoría de los casos eso no era así[5]. Estos estudios críticos se siguen citando ampliamente por quienes sustituirían a todos los expertos en un tema (p. ej., cardiólogos) por expertos en búsqueda y evaluación (personas especializadas en buscar

Cuadro 9.1 Ventajas de las revisiones sistemáticas[2]

- Los métodos explícitos *limitan el sesgo* a la hora de identificar y rechazar los estudios.
- Las conclusiones son, por lo tanto, más *fiables* y *precisas.*
- Se pueden asimilar grandes cantidades de *información* rápidamente por parte de los profesionales sanitarios, investigadores y elaboradores de políticas.
- El retraso entre los descubrimientos de la investigación y la *aplicación* de las estrategias diagnósticas y terapéuticas eficaces se reduce (v. cap. 12).
- Los resultados de los diferentes estudios pueden compararse formalmente para establecer la *generalizabilidad* de los resultados y su *concordancia* (falta de heterogeneidad) (v. sección «Cocientes de verosimilitudes»).
- Se pueden identificar las razones de la *heterogeneidad* (discordancia entre los resultados de los estudios) y es posible generar nuevas hipótesis sobre subgrupos particulares (v. sección «Cocientes de verosimilitudes»).
- La revisiones sistemáticas cuantitativas (metaanálisis) aumentan la *precisión* del resultado global (v. secciones «¿Se abordaron las cuestiones estadísticas preliminares?» y «Diez preguntas que deben plantearse sobre un artículo que pretende validar una prueba diagnóstica o de cribado»).

y criticar artículos sobre cualquier tema). Sin embargo, en los últimos años nadie ha vuelto a observar unos hallazgos similares. Dicho de otro modo, tal vez deberíamos confiar en que los expertos actuales sean más tendentes a basar sus recomendaciones en una evaluación exhaustiva de la evidencia. Como regla general, si se quiere buscar la mejor evidencia objetiva sobre los beneficios de (por ejemplo) diferentes anticoagulantes sobre la fibrilación auricular, se debería pedir a un experto en revisiones sistemáticas que colabore con un experto en fibrilación auricular.

Para ser justos con Pauling[2], él mencionó una serie de ensayos cuyos resultados cuestionaban seriamente su teoría de que la vitamina C previene el resfriado común, pero atribuyó «defectos metodológicos» a todos ellos. Muchos de los ensayos que Pauling *sí* incluyó en su análisis también tenían dichos defectos, pero como sus resultados concordaban con la opinión de Pauling, él fue, tal vez inconscientemente, menos crítico con los puntos débiles de su diseño[6].

Pongo este ejemplo para ilustrar el hecho de que, cuando se realiza una revisión sistemática, además de efectuar una búsqueda exhaustiva y objetiva de los artículos pertinentes, para rechazar los artículos «defectuosos» se deben aplicar unos criterios explícitos e independientes de los resultados de esos ensayos. Dicho de otro modo, no se debe evaluar críticamente un ensayo porque todos los demás ensayos sobre el tema muestren algo diferente (v. sección «Explicación

de la heterogeneidad»); dicha evaluación crítica debe realizarse porque, *con independencia de los resultados que se presenten,* los objetivos o métodos del ensayo no cumplieron nuestros criterios de inclusión o estándares de calidad (v. sección «La ciencia de criticar los artículos»).

Evaluación de las revisiones sistemáticas

Uno de los principales avances en la medicina basada en la evidencia (MBE) desde que escribí la primera edición de este libro en 1995 ha sido el consenso sobre un formato estándar y estructurado para escribir y publicar las revisiones sistemáticas, cuya versión original se denominó *declaración QUORUM* (equivalente al formato CONSORT para la publicación de ensayos controlados aleatorizados descrito en la sección «Ensayos controlados aleatorizados»). Dicho formato se actualizó posteriormente y pasó a denominarse declaración PRISMA (acrónimo de *Preferred Reporting Items for Systematic Reviews and Meta-Analyses,* es decir, elementos de información preferidos para revisiones sistemáticas y metaanálisis)[7]. Si se siguen estas listas de comprobación estructuradas, las revisiones sistemáticas y metaanálisis se convierten en unos elementos mucho más fáciles de usar a la hora de buscar información. A continuación se presentan algunas preguntas basadas en la lista de comprobación PRISMA (pero muy resumidas y simplificadas) que se pueden plantear acerca de cualquier revisión sistemática de la evidencia cuantitativa.

Pregunta uno: ¿Cuál es la pregunta clínica importante que aborda la revisión?

En el capítulo 3 se explicó la importancia de definir la pregunta al leer un artículo acerca de un ensayo clínico o de otro tipo de investigación primaria. El capítulo se titulaba «*Aprendiendo a orientarse*» porque si no se logra determinar sobre qué trata un artículo, su lectura es una forma segura de acabar más confundido. La definición de una pregunta específica que pueda contestarse es aún más importante (y a menudo se omite con demasiada frecuencia) a la hora de preparar una revisión de los estudios primarios. Quien haya tratado de recopilar alguna vez las conclusiones de varios artículos clínicos en un trabajo de revisión, editorial o notas resumidas para un examen sabrá que es demasiado fácil perderse en aspectos del tema que no estaba previsto abarcar.

La cuestión abordada por una revisión sistemática se debe definir de forma muy precisa, pues el revisor debe tomar una decisión dicotómica (sí/no) sobre si cada artículo que pueda tener relevancia se incluirá o, por el contrario, se rechazará por «irrelevante». La pregunta: «¿los anticoagulantes previenen los accidentes cerebrovasculares (ictus) en pacientes con fibrilación auricular?» parece bastante específica hasta que se empieza a mirar la lista de posibles estudios que se van a incluir. ¿Incluye la «fibrilación auricular» las formas tanto reumáticas como no reumáticas (que se sabe que están asociadas con riesgos muy diferentes de ictus), e incluye la fibrilación auricular intermitente? Mi abuelo, por ejemplo, solía desarrollar esta arritmia durante unas horas en las pocas ocasiones en que bebía café y se habría contado como un «caso gris» en cualquier ensayo.

¿Incluye el «ictus» tanto las formas isquémicas (causadas por una *obstrucción* de un vaso sanguíneo cerebral) como hemorrágicas (causadas por una *rotura* de un vaso sanguíneo)? Además, hablando de vasos sanguíneos rotos, ¿no deberían sopesarse los efectos secundarios de los anticoagulantes frente a sus posibles beneficios? ¿Se utiliza «anticoagulante» en el sentido estricto del término (es decir, fármacos que actúan en la cascada de la coagulación), como la heparina, warfarina y dabigatrán, o también se incluyen fármacos que reducen la agregación plaquetaria, como la aspirina y el clopidogrel? Por último, ¿la revisión debería incluir los ensayos sobre las personas que ya han sufrido un ictus o un ataque isquémico transitorio (un ictus leve que mejora en 24 h), o debe limitarse a los ensayos en personas sin estos factores de riesgo significativos para sufrir otro ictus? La pregunta «sencilla» planteada anteriormente empieza a convertirse en una sin respuesta y debe concretarse del siguiente modo:

> *Evaluación de la eficacia y seguridad del tratamiento anticoagulante de tipo warfarina en la prevención secundaria (es decir, después de un ictus o ataque isquémico transitorio previo) en pacientes con cualquier forma de fibrilación auricular: comparación con el tratamiento antiagregante plaquetario*[8].

Pregunta dos: ¿Se ha realizado una búsqueda exhaustiva en la(s) base(s) de datos apropiada(s) y se han explorado otras fuentes potencialmente importantes?
Como se ilustra en la figura 9.1, uno de los beneficios de una revisión sistemática es que, a diferencia de una revisión narrativa o periodística, se requiere que el autor indique de dónde procede la información y cómo se ha procesado. Como se explicó en el capítulo 2, la búsqueda de los artículos pertinentes en la base de datos Medline es una ciencia sofisticada e incluso la mejor búsqueda en Medline no incluirá algunos artículos importantes. Un revisor que busque un conjunto exhaustivo de estudios primarios debe consultar las otras bases de datos que figuran en la sección «Estudios primarios: desentrañando la selva», y a veces muchos más (p. ej., en una revisión sistemática de la difusión de innovaciones en las organizaciones de servicios sanitarios, mis colegas y yo buscamos en un total de 15 bases de datos, de 9 de las cuales nunca había oído hablar cuando empecé el estudio[9]).

En la búsqueda de los ensayos que se deben incluir en una revisión, es obligatorio desde los puntos de vista científico y político evitar de forma escrupulosa el imperialismo lingüístico. Se debe conceder el mismo peso a las expresiones «Eine Placebo-kontrollierte Doppel-blindstudie» y «une étude randomisée a double insu face au placebo» como a «a double-blind, randomised controlled trial» y a «ensayo doble ciego, aleatorizado y controlado»[6], aunque la omisión de estudios de otro idioma no suele asociarse a resultados sesgados (simplemente es mala ciencia)[10]. Además, sobre todo cuando se está considerando realizar una síntesis estadística de los resultados (metaanálisis), puede ser necesario escribir y preguntar a los autores de los estudios primarios por los datos que no

se incluyeron originalmente en la revisión publicada (v. sección «Metaanálisis para no estadísticos»).

Incluso cuando se ha realizado todo esto, la búsqueda sistemática de material por parte del revisor no ha hecho más que empezar. Como Knipschild y cols.[6] demostraron cuando realizaron una búsqueda de ensayos sobre la vitamina C y la prevención del resfriado, sus bases de datos electrónicas sólo les proporcionaron 22 de su número total definitivo de 61 ensayos. Otros 39 ensayos se descubrieron mediante una búsqueda manual de la base de datos Index Medicus (14 ensayos no identificados previamente) y buscando las referencias de los ensayos identificados en Medline (15 ensayos más), las referencias de las referencias (otros 9 ensayos) y las referencias de las referencias de las referencias (un ensayo adicional no identificado en ninguna de las búsquedas anteriores). Sin embargo, no hay que ser demasiado duro con un revisor si no ha seguido este consejo de perfección al pie de la letra. Después de todo, Knipschild y cols.[6] observaron que sólo uno de los ensayos no identificados en Medline cumplía con los criterios estrictos de calidad metodológica y en última instancia contribuyó a su revisión sistemática sobre la vitamina C en la prevención del resfriado. El uso de métodos de búsqueda más laboriosos (como la búsqueda de las referencias de referencias, encadenamiento de citas escribiendo a todos los expertos conocidos en la materia, y la búsqueda de «literatura gris») (v. cuadro 9.2 y también la sección «Estudios primarios: desentrañando la selva») puede tener mayor importancia relativa cuando se analizan ensayos que no están en la literatura médica principal. Por ejemplo, en la gestión de los servicios sanitarios, mi propio equipo demostró que sólo alrededor del 25% de los artículos de alta calidad relevantes se obtuvieron mediante una búsqueda electrónica[11].

Cuadro 9.2 Lista de comprobación de las fuentes de datos para una revisión sistemática

- Base de datos Medline.
- Registro de ensayos clínicos controlados Cochrane (v. «Fuentes sintetizadas», pág. 17).
- Otras bases de datos médicas y paramédicas (v. todo el cap. 2, pág. 15).
- Literatura en otros idiomas.
- «Literatura gris» (tesis, informes internos, revistas sin revisión por pares, archivos de la industria farmacéutica).
- Referencias (y referencias de referencias, etc.) que figuran en las fuentes primarias.
- Otras fuentes no publicadas conocidas por los expertos en la materia (búsqueda mediante comunicación personal).
- Datos en bruto de los ensayos publicados (búsqueda por comunicación personal).

Pregunta tres: ¿Se evaluó la calidad metodológica y se ponderaron los ensayos en consonancia?

En los capítulos 3 y 4 y en el apéndice 1 se proporcionan algunas listas de comprobación para evaluar si un artículo debe ser rechazado de plano por razones metodológicas. Sin embargo, teniendo en cuenta que sólo alrededor del 1% de los ensayos clínicos se consideran libres de críticas metodológicas, la cuestión práctica es cómo garantizar que un estudio «pequeño pero perfectamente realizado» reciba el peso que merece respecto a un estudio más amplio cuyos métodos son adecuados, pero más abiertos a la crítica. Como la declaración PRISMA subraya, la pregunta clave es en qué grado es probable que los defectos metodológicos hayan *sesgado* los resultados de la revisión[7].

Las deficiencias metodológicas que invalidan los resultados de los ensayos suelen ser genéricas (es decir, son independientes del tema de estudio; v. apéndice 1), pero también puede haber ciertas características metodológicas que distingan entre una calidad buena, intermedia y baja en un campo particular. Por lo tanto, una de las tareas de un revisor sistemático es la elaboración de una lista de criterios, incluidos los aspectos tanto genéricos como particulares de calidad, respecto a los que evaluar cada ensayo. En teoría, se debería calcular una puntuación numérica compuesta que refleje la «calidad metodológica global». En realidad, sin embargo, se debe tener cuidado al elaborar estas puntuaciones, pues no existe un patrón oro de la «verdadera» calidad metodológica de un ensayo y estas puntuaciones compuestas pueden no ser válidas ni fiables en la práctica. Quien tenga interés en leer más sobre la ciencia de la elaboración y aplicación de criterios de calidad a los estudios como parte de una revisión sistemática puede consultar la última edición del *Manual Cochrane de revisores*[12].

Pregunta cuatro: ¿Qué grado de sensibilidad tienen los resultados respecto al modo en que se ha realizado la revisión?

Quien no entienda lo que significa esta pregunta debe leer el artículo irónico publicado por Counsell y cols.[13] hace algunos años en el *British Medical Journal,* que «demostraba» una relación totalmente espuria entre el resultado de lanzar un dado y el resultado de un accidente cerebrovascular agudo. Los autores describían una serie de experimentos artificiales sobre lanzar unos dados, en los que unos dados de color rojo, blanco y verde, respectivamente, representaban diferentes tratamientos del accidente cerebrovascular agudo.

En general, los «ensayos» mostraron que ninguna de las tres terapias tenía beneficios significativos. Sin embargo, la simulación de una serie de eventos perfectamente plausibles en el proceso de metaanálisis, como la exclusión de varios de los ensayos «negativos» debido al sesgo de publicación (v. sección «Ensayos controlados aleatorizados»), un análisis de subgrupos que excluyó los datos sobre el tratamiento de los dados rojos (porque, al volver a evaluar los resultados, los dados rojos parecían ser perjudiciales) y otras exclusiones básicamente arbitrarias por motivos de «calidad metodológica», permitió concluir que existía un beneficio aparentemente muy significativo del «tratamiento con dados» en el ictus agudo.

Es evidente que no se puede curar a nadie de un ictus lanzando un dado, pero si estos resultados simulados hubiesen correspondido a una auténtica controversia médica (p. ej., si sería mejor aconsejar a las mujeres posmenopáusicas que tomasen un tratamiento hormonal sustitutivo o si los bebés que vienen de nalgas deberían extraerse sistemáticamente por cesárea), ¿cómo se podrían detectar estos sesgos sutiles? La respuesta es que hay que moverse en el terreno de lo hipotético. ¿Qué pasaría si los autores de la revisión sistemática hubiesen cambiado los criterios de inclusión? ¿Qué pasaría si hubiesen excluido los estudios no publicados? ¿Qué pasaría si sus «ponderaciones de calidad» se hubiesen asignado de forma diferente? ¿Qué pasaría si se hubieran incluido (o excluido) ensayos de menor calidad metodológica? ¿Qué pasaría si se asumiese que todos los pacientes no contabilizados de un ensayo hubiesen fallecido (o se hubiesen curado)?

Un análisis de estos planteamientos hipotéticos se denomina *análisis de sensibilidad*. Si resulta que manipular con los datos de este tipo de diversas maneras provoca unas variaciones escasas o nulas en los resultados generales de la revisión, se puede asumir que las conclusiones de la misma son relativamente sólidas. Sin embargo, si los principales hallazgos desaparecen cuando cualquiera de las posibilidades hipotéticas varía, las conclusiones deberían expresarse con mucha más cautela y habría que dudar antes de modificar las prácticas basándose en ellas.

Pregunta cinco: ¿Se han interpretado los resultados numéricos con sentido común y con la debida atención a los aspectos más generales del problema?

Como se muestra en la siguiente sección, es fácil verse abrumado por las cifras y gráficos de una revisión sistemática. Sin embargo, cualquier resultado numérico, aunque sea preciso, exacto, significativo o incontrovertible, se debe situar en el contexto de la extremadamente simple y (a menudo) frustrante pregunta general planteada por la revisión. El médico debe decidir de qué modo (si se da el caso) este resultado numérico, *ya sea significativo o no,* debería influir en la asistencia de un paciente individual.

Una característica fundamental que se debe tener en cuenta al llevar a cabo o evaluar una revisión sistemática es la validez externa de los ensayos incluidos (v. cuadro 9.3). Un ensayo puede tener una calidad metodológica elevada y ofrecer un resultado preciso y numéricamente impresionante, pero puede, por ejemplo, haberse realizado con participantes menores de 60 años, por lo que tal vez no sea aplicable a las personas mayores de 75 años por razones fisiológicas evidentes. La inclusión de estudios irrelevantes en las revisiones sistemáticas es una garantía para llegar a conclusiones absurdas y reducir la credibilidad de la investigación secundaria.

Metaanálisis para no estadísticos

Si tuviera que elegir una palabra que ejemplificase el miedo y el odio que sienten muchos estudiantes, médicos y consumidores hacia la MBE, esa palabra sería «metaanálisis». El metaanálisis, que se define como *una síntesis estadística de*

Cuadro 9.3 Ponderación de los ensayos en una revisión sistemática

Cada ensayo debe ser evaluado en términos de su:

- *Calidad metodológica,* es decir, hasta qué punto es probable que el diseño y la realización hayan evitado los errores sistemáticos (sesgo) (v. sección «¿Se evitó o se minimizó el sesgo sistemático?»).
- *Precisión,* que constituye una medida de la probabilidad de errores aleatorios (representada habitualmente como la anchura del intervalo de confianza alrededor del resultado).
- *Validez externa,* es decir, el grado en que los resultados son generalizables o aplicables a una población diana particular.

(Los revisores y editores de revistas conceden, con razón, un gran peso a otros aspectos adicionales de «calidad», como la importancia científica, importancia clínica y calidad literaria, pero son menos relevantes para el revisor sistemático una vez que se ha definido la pregunta que se va a abordar.)

los resultados numéricos de varios ensayos que abordan la misma pregunta, es la oportunidad que tienen los estadísticos de asestarnos un doble golpe. En primer lugar, nos asustan con todas las pruebas estadísticas de los artículos individuales y luego usan una batería completamente nueva de pruebas para presentar una nueva serie de odds ratio, intervalos de confianza y valores de significación.

Como ya expuse en el capítulo 5, yo también tiendo a dejarme llevar por el pánico ante la visión de las proporciones, los signos de raíz cuadrada y las letras griegas casi olvidadas, pero antes de encasillar los metaanálisis en el conjunto de técnicas especializadas que nunca llegaremos a comprender, hay que recordar dos cosas. En primer lugar, puede que el metaanalista viva en su mundo lejano de la estadística, pero está *de nuestra parte.* Un buen metaanálisis suele ser más fácil de entender para los no estadísticos que el conjunto de artículos de investigación primarios a partir de los que se ha elaborado, por las razones que se explicarán a continuación. En segundo lugar, los principios estadísticos subyacentes utilizados para el metaanálisis son los mismos que los de cualquier otro análisis de datos. La única diferencia es que algunos de los números son más grandes.

La primera tarea del metaanalista, después de seguir los pasos preliminares para la revisión sistemática de la figura 9.1, es decidir cuál o cuáles de las diversas medidas de resultado elegidas por los autores de los estudios primarios son las mejores para usarlas en la síntesis global. En los ensayos sobre un régimen de quimioterapia en particular para el cáncer de mama, por ejemplo, algunos autores publicarán las cifras de mortalidad acumuladas (es decir, el número total de personas que han fallecido hasta la fecha) en los puntos de corte de 3 y 12 meses, mientras que en otros ensayos se publicará la mortalidad acumulada a los 6 meses, 12 meses y 5 años. El metaanalista puede decidir concentrarse en la mortalidad a los 12 meses, ya que este resultado se puede obtener fácilmente en todos los artículos. No obstante, puede decidir que la mortalidad a los 3 meses

es un criterio de valoración clínicamente importante y tendría que escribir a los autores de los ensayos restantes para solicitarles los datos en bruto a partir de los cuales se calcularon estas cifras.

Además de procesar los números, parte de la descripción del trabajo del metaanalista consiste en tabular la información relevante sobre los criterios de inclusión, tamaño muestral, características basales de los pacientes, tasa de abandono («exclusiones») y resultados de los criterios de valoración principales y secundarios de todos los estudios incluidos. Si esta tarea se ha realizado correctamente, será posible comparar tanto los métodos como los resultados de dos ensayos cuyos autores hayan redactado sus investigaciones de forma diferente. Aunque estas tablas suelen ser visualmente abrumadoras, ahorran tener que bucear en las secciones de métodos de cada artículo y comparar los resultados tabulados de un autor con el gráfico circular o histograma de otro autor.

En la actualidad, los resultados de los metaanálisis tienden a presentarse de forma bastante estándar. Esto se debe en parte a que los metaanalistas suelen utilizar programas informáticos para hacer los cálculos (véase la última edición del *Manual Cochrane de revisores* para consultar un menú actualizado de opciones[12]) y la mayoría de los paquetes de software incluyen una herramienta de gráficos estándar que presenta los resultados como se muestra en la figura 9.2. Se ha reproducido (con permiso de los autores) esta representación gráfica (denominada coloquialmente *diagrama* o *gráfico de bosque*) de las odds ratio combinadas de ocho ensayos controlados aleatorizados sobre el tratamiento de la depresión. En cada uno de estos ocho estudios se había comparado un grupo que recibió terapia cognitivo-conductual (TCC) frente a un grupo control que no recibió el

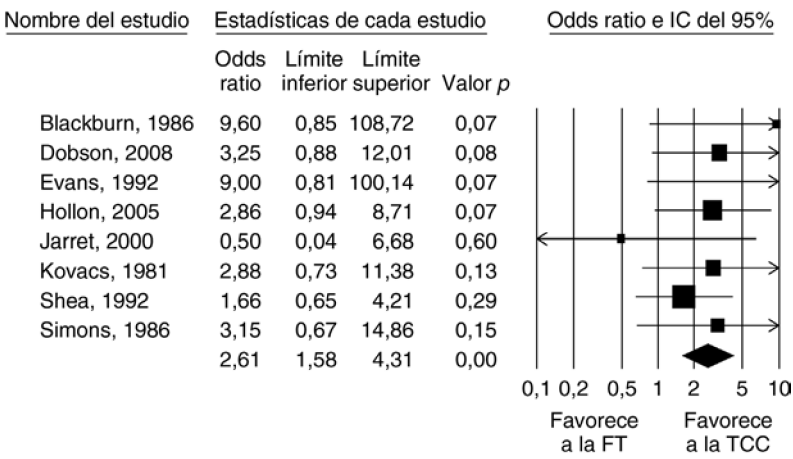

Figura 9.2 Diagrama de bosque que muestra los efectos a largo plazo de la terapia cognitivo-conductual (TCC) en comparación con ningún tratamiento activo y la interrupción de la farmacoterapia (FT). Fuente: Cuijpers y cols.[14]. Reproducida con autorización del BMJ.

tratamiento activo y en el que la farmacoterapia (FT, es decir, tratamiento con medicamentos) se suspendió[14]. El resultado primario (principal) de este metaanálisis fue la recidiva antes de 1 año.

Los ocho ensayos, cada uno representado por el apellido del primer autor y el año de publicación del artículo (p. ej., «Blackburn, 1986») se presentan uno debajo del otro en el lado izquierdo de la figura. La línea horizontal correspondiente a cada ensayo muestra la probabilidad de recidiva al año en los pacientes asignados de forma aleatoria a la TCC en comparación con los pacientes asignados aleatoriamente a la FT. La «mancha» en el centro de cada línea es la estimación puntual de la diferencia entre los grupos (la mejor estimación individual del beneficio a la hora de mejorar la tasa de recidiva si se utiliza TCC en lugar de FT) y la anchura de la línea representa el intervalo de confianza del 95% de esta estimación (v. sección «¿Se han calculado los intervalos de confianza, y se reflejan en las conclusiones de los autores?»). La línea vertical que es fundamental observar, denominada *línea de ausencia de efecto,* es la que marca el riesgo relativo (RR) de 1,0. Hay que señalar que, si la línea horizontal de cualquier ensayo no cruza la línea de ausencia de efecto, hay un 95% de probabilidades de que haya una diferencia «real» entre los grupos.

Como se ha expuesto en las secciones «¿Se abordaron cuestiones estadísticas preliminares?» y «Probabilidad y confianza», si el intervalo de confianza del resultado (la línea horizontal) *cruza* la línea de ausencia de efecto (es decir, la línea vertical donde RR = 1,0), esto puede reflejar que no hay diferencias significativas entre los tratamientos *y/o* que el tamaño de la muestra era demasiado pequeño para distinguir con confianza entre lo verdadero y lo falso. Los diversos estudios individuales ofrecen estimaciones puntuales de la odds ratio de la TCC en comparación con la FT (de entre 0,5 y 9,6), y los intervalos de confianza de algunos estudios son tan amplios que incluso se salen del gráfico.

Ahora viene lo divertido de los metaanálisis. Prestemos atención al pequeño rombo situado debajo de todas las líneas horizontales. Representa los datos *combinados* de los ocho ensayos (RR global TCC:FT = 2,61, lo que significa que la TCC tiene 2,61 veces más posibilidades de prevenir la recidiva), con un nuevo intervalo de confianza, mucho más estrecho, de este RR (1,58-4,31). Puesto que el rombo no se superpone a la línea de ausencia de efecto, es posible afirmar que hay una diferencia estadísticamente significativa entre los dos tratamientos en términos del criterio de valoración principal (recidiva de la depresión en el primer año). No obstante, en este ejemplo, siete de los ocho ensayos sugirieron un beneficio de la TCC, pero en ninguno de ellos el tamaño de la muestra era lo bastante grande como para que la conclusión fuese estadísticamente significativa.

Sin embargo, se debe tener en cuenta que este pequeño y nítido rombo *no* significa que se deba ofrecer la TCC a todos los pacientes con depresión. Su significado es mucho más limitado: es probable que el paciente *promedio* de los ensayos recogidos en este metaanálisis se beneficie en términos del resultado primario (recidiva de la depresión en el plazo de un año) si recibe TCC. La elección

Figura 9.3 Logotipo de la Colaboración Cochrane.

del tratamiento debería tener en cuenta la opinión del paciente acerca de recibir TCC (v. cap. 16) y también los beneficios relativos de esta terapia en comparación con *otros* tratamientos para la depresión. En el artículo del que se ha tomado la figura 9.2 también se describía un segundo metaanálisis que no mostró diferencias significativas entre la TCC y la continuación del tratamiento antidepresivo, lo que sugiere, tal vez, que los pacientes que prefieren no recibir TCC pueden evolucionar igual de bien si continúan tomando su medicación[14].

Como muestra este ejemplo, los ensayos «no significativos» (es decir, los que, por sí solos, no mostraron una diferencia significativa entre los grupos de tratamiento y control) contribuyen a un resultado combinado en un metaanálisis que es estadísticamente significativo. El ejemplo más famoso de esto, que la Colaboración Cochrane adoptó como su logotipo (fig. 9.3), es el metaanálisis de siete ensayos sobre el efecto de administrar esteroides a las madres en quienes se preveía un parto prematuro[15]. Sólo dos de los siete ensayos mostraron un beneficio estadísticamente significativo (en términos de supervivencia del lactante), pero la mejora de la precisión (es decir, el estrechamiento de los intervalos de confianza) en los resultados combinados, que se muestra por la anchura más estrecha del rombo en comparación con las líneas individuales, demuestra la fuerza de la evidencia a favor de esta intervención. Este metaanálisis puso de manifiesto que los lactantes de las madres tratadas con esteroides tenían un 30-50% menos de probabilidades de fallecer que los de las madres del grupo control. Este ejemplo se comentará con más detalle en la sección «¿Por qué los profesionales sanitarios adoptan lentamente la práctica basada en la evidencia?» en relación con el cambio del comportamiento de los médicos.

Llegados a este punto, es posible que el lector haya caído en la cuenta de que cualquiera que esté pensando llevar a cabo un ensayo clínico de una intervención primero debería llevar a cabo un metaanálisis de todos los ensayos anteriores sobre la misma intervención. En la práctica, los investigadores sólo hacen esto de vez en cuando. Dean Fergusson y cols., del Health Research Institute de Ottawa, publicaron un metaanálisis acumulativo de todos los ensayos controlados aleatorizados realizados sobre el uso del fármaco aprotinina para la hemorragia perioperatoria durante la cirugía cardíaca[16]. Organizaron los ensayos en el orden en que se habían publicado y observaron lo que un metaanálisis de «todos los ensayos realizados hasta la fecha» habría mostrado (si se hubiera realizado en ese momento). El *metaanálisis acumulativo* resultante ofreció unos resultados impactantes para la comunidad de investigadores. El efecto beneficioso de la aprotinina alcanzó significación estadística después de sólo 12 ensayos, es decir, en 1992. Sin embargo, dado que nadie hizo un metaanálisis en ese momento, se llevaron a cabo otros 52 ensayos clínicos (y es posible que se estén efectuando más todavía). Todos estos ensayos fueron innecesarios desde el punto de vista científico y poco éticos (porque la mitad de los pacientes no recibió un fármaco del que ya se había demostrado que mejoraba el resultado). En la figura 9.4 se muestra este derroche de esfuerzo.

Los lectores que hayan seguido hasta aquí la argumentación sobre el metaanálisis de los resultados de los ensayos publicados pueden leer una descripción de la técnica más sofisticada del metaanálisis de datos de pacientes concretos, que proporciona una cifra más exacta y precisa de la estimación puntual del efecto[17]. También pueden consultar el libro de texto que se está convirtiendo en un clásico sobre el tema[18].

Explicación de la heterogeneidad

En el lenguaje cotidiano, «homogéneo» significa «de composición uniforme», y «heterogéneo» «muchos ingredientes diferentes». En el lenguaje del metaanálisis, la homogeneidad significa que los resultados de cada ensayo individual son compatibles con los resultados de cualquiera de los otros ensayos. La homogeneidad se puede estimar a simple vista cuando los resultados de los ensayos se presentan en el formato que se muestra en las figuras 9.2 y 9.5. En la figura 9.2, el límite inferior del intervalo de confianza de cada ensayo está por debajo del límite superior del intervalo de confianza de todos los demás (es decir, todas las líneas horizontales se superponen en cierta medida). Desde el punto de vista estadístico, los ensayos son homogéneos. En cambio, en la figura 9.4 hay algunos ensayos cuyos límites inferiores del intervalo de confianza están por encima del límite superior del intervalo de confianza de uno o más de otros ensayos (es decir, algunas líneas no se superponen en absoluto). Estos ensayos se pueden considerar heterogéneos.

Llegados a este punto, es posible que el lector haya advertido (sobre todo si ya ha leído la sección «¿Se han calculado los intervalos de confianza, y se reflejan en las conclusiones de los autores?» sobre los intervalos de confianza) que calificar como

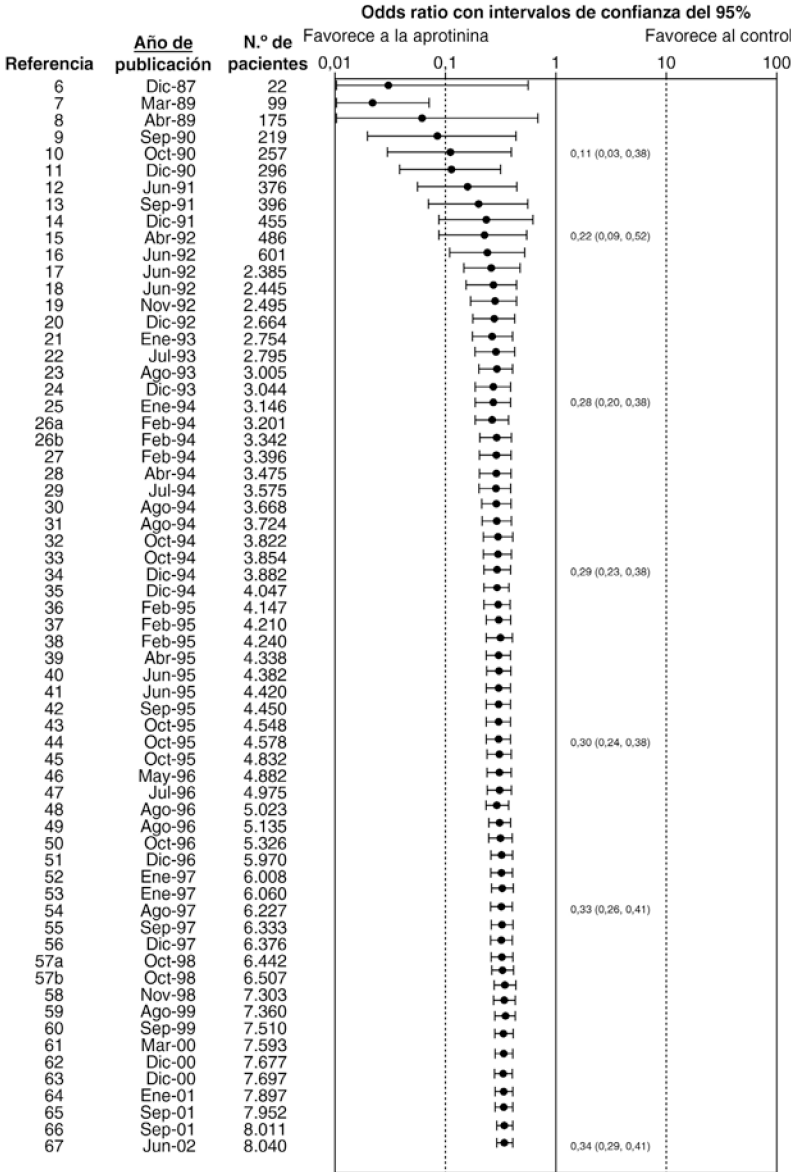

Figura 9.4 Metaanálisis acumulativos de ensayos controlados aleatorizados sobre la aprotinina en la cirugía cardíaca[16]. Reproducida con autorización de Clinical Trials.

heterogénea una serie de ensayos en función de si sus intervalos de confianza se superponen es algo arbitrario, como también lo es el propio intervalo de confianza (se puede fijar en el 90%, 95%, 99% o cualquier otro valor). La prueba definitiva implica utilizar una herramienta estadística ligeramente más sofisticada que colocar una regla sobre un diagrama de bosque. La más usada es una variante de la prueba de chi cuadrado (χ^2) (v. tabla 5.1), ya que la cuestión abordada es «¿hay mayor variación entre los resultados de los ensayos de lo que es compatible con el azar?».

El estadístico χ^2 para la heterogeneidad se explica con más detalle en la publicación de Thompson[19], que ofrece la siguiente regla general: un estadístico χ^2 tiene, en promedio, un valor igual a sus grados de libertad (en este caso, el número de ensayos en el metaanálisis menos uno), por lo que un χ^2 de 7,0 para un conjunto de 8 ensayos no proporcionaría evidencia de heterogeneidad estadística. (De hecho, tampoco demostraría que los ensayos eran homogéneos, sobre todo porque la prueba de χ^2 tiene una potencia baja [v. sección «¿Se abordaron las cuestiones estadísticas preliminares?»] para detectar niveles de heterogeneidad pequeños pero importantes.)

Un valor χ^2 mucho mayor que el número de ensayos en un metaanálisis indica que los ensayos que contribuyeron al análisis presentan alguna diferencia importante entre sí. Puede haber, por ejemplo, diferencias metodológicas conocidas (p. ej., los autores pueden haber utilizado distintos cuestionarios para evaluar los síntomas de la depresión) o diferencias clínicas conocidas entre los participantes del ensayo (p. ej., un centro podría haber sido un hospital terciario de referencia al cual se derivasen todos los pacientes más enfermos). Sin embargo, puede haber diferencias desconocidas o no registradas entre los ensayos sobre las cuales el metaanalista sólo puede especular hasta haber obtenido más detalles de los autores de los ensayos. Se debe recordar que demostrar la heterogeneidad estadística es un ejercicio matemático y es una tarea del estadístico, pero *la explicación* de esta heterogeneidad (es decir, buscar y explicar la heterogeneidad *clínica*) es un ejercicio interpretativo y requiere imaginación, sentido común y práctica clínica o experiencia en investigación.

En la figura 9.5, que se reproduce con autorización del capítulo de Thompson[19] sobre el tema, se muestran los resultados de diez ensayos sobre estrategias para reducir el colesterol. Los resultados se expresan como la reducción porcentual del riesgo de cardiopatía asociado con cada disminución de 0,6 mmol/l del nivel sérico de colesterol. Las líneas horizontales representan los intervalos de confianza del 95% de cada resultado y es evidente, incluso sin conocer que el estadístico χ^2 es de 127, que los ensayos son muy heterogéneos.

Realizar simplemente una «media» de los resultados de los ensayos de la figura 9.5 sería muy engañoso. El metaanalista debe volver a sus fuentes primarias y preguntar: «¿en qué difería el ensayo A del ensayo B y qué tienen en común los ensayos E, F y H que hace que sus resultados se agrupen en un extremo de la figura?». En este ejemplo, una corrección para la edad de los participantes en el ensayo redujo el valor de χ^2 de 127 a 45. Dicho de otro modo, la mayor parte de la «incompatibilidad» en los resultados de estos ensayos se puede explicar por

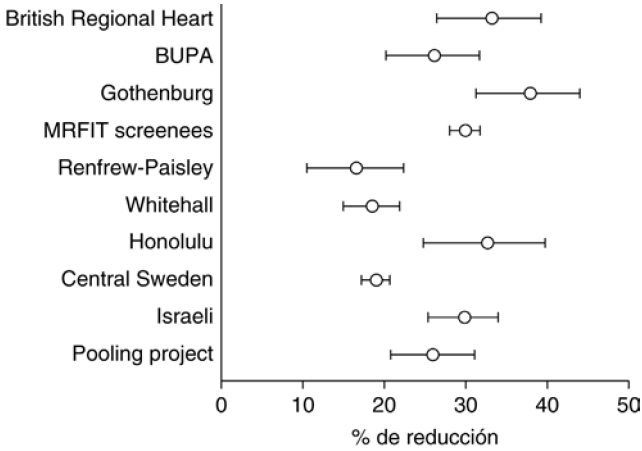

Figura 9.5 Reducción del riesgo de cardiopatía mediante las estrategias para reducir el colesterol. Fuente: Greenhalgh[20]. Reproducida con autorización del Royal College of General Practitioners.

el hecho de que aplicar una estrategia (como una dieta especial) que reduzca adecuadamente el nivel de colesterol tendrá muchas más probabilidades de evitar un ataque cardíaco en las personas de 45 años que en las de 85.

Esto constituye la queja fundamental del profesor Hans Eysenck[21], quien ha elaborado una crítica enérgica y entretenida de la ciencia del metaanálisis. Si dividiésemos a las personas entre las que se centran en las similitudes y las que prestan más atención a las diferencias, Eysenck estaría en el segundo grupo, y la combinación de los resultados de los estudios que se realizaron en diferentes poblaciones, en distintos lugares, en diferentes momentos y por distintas razones ofende su sentido de lo cualitativo y lo particular (v. cap. 12).

Las reservas de Eysenck sobre los metaanálisis quedaron reflejadas en el metaanálisis tristemente desprestigiado que demostró (erróneamente) que se podía obtener un beneficio significativo al administrar magnesio intravenoso a las víctimas de ataques cardíacos. En un megaensayo posterior que incluyó a 58.000 pacientes (ISIS-4) no se observó ningún beneficio, y las conclusiones engañosas de los metaanalistas se explicaron posteriormente en términos de sesgo de publicación, debilidades metodológicas de los ensayos más pequeños y heterogeneidad clínica[22,23]. (Por cierto, se puede leer más sobre el debate sobre los pros y los contras de los metaanálisis frente a los megaensayos en una publicación reciente[24].)

La ingenuidad matemática de Eysenck resulta embarazosa («si un tratamiento médico tuviese un efecto tan recóndito y oscuro como para requerir un metaanálisis para demostrarlo, no me gustaría que lo usasen en mí»), lo que quizá explica por qué los editores de la segunda edición del libro *Systematic reviews* eliminaron su capítulo de la colección. Sin embargo, yo siento gran simpatía por los fundamentos de su argumentación. Dado que tiendo a alinearme con las personas

que prestan más atención a las diferencias, pondría los recelos de Eysenck sobre los metaanálisis en un puesto destacado de la lista de lecturas obligatorias para los aspirantes a revisor sistemático. De hecho, una vez yo misma participé en el debate al colaborar con Griffin[25] en la publicación de un metaanálisis de los estudios primarios sobre el tratamiento de la diabetes por los equipos de atención primaria. Aunque siento gran respeto por Simon como científico, creía seriamente que él no tenía razones para realizar una suma matemática de lo que, en mi opinión, eran estudios muy diferentes que abordaban preguntas ligeramente distintas. Como dije en mi comentario sobre su artículo, «cuatro manzanas y cinco naranjas suman cuatro manzanas y cinco naranjas, no nueve manzanas y naranjas»[26]. Sin embargo, Simon se considera a sí mismo un autor que presta más atención a las similitudes y hay mucha gente más inteligente que yo que ha argumentado que estaba totalmente en lo cierto al analizar sus datos como lo hizo. Por fortuna, ambos hemos llegado al acuerdo de disentir y seguimos siendo amigos a nivel personal.

Nuevos enfoques de la revisión sistemática

En este capítulo se ha descrito el enfoque más utilizado para las revisiones sistemáticas: la síntesis de los ensayos terapéuticos. Quien crea que ha comprendido todo lo expuesto, tal vez quiera comenzar a explorar la literatura sobre los tipos más difíciles de revisión sistemática, como los estudios de diagnóstico[27] y la ciencia emergente de la revisión sistemática de la investigación cualitativa (y los estudios mixtos cualitativos y cuantitativos), que se describen con más detalle en el capítulo 11. Por mi parte, he trabajado con algunos colegas para desarrollar nuevos enfoques de la revisión sistemática que subrayan y analizan (en lugar de intentar «promediar») las diferencias fundamentales entre los estudios primarios, un enfoque que considero especialmente útil para elaborar revisiones sistemáticas en la elaboración de políticas sanitarias[28,29]. Sin embargo, estas aplicaciones relativamente menores no son algo básico y quien esté leyendo este libro para preparar un examen probablemente verá que no están en el plan de estudios.

Quien haya simpatizado con los planteamientos del profesor Eysenck expuestos en el apartado anterior, puede consultar otras críticas teóricas sobre la revisión sistemática. MacLure[30] ha escrito un artículo filosófico excelente afirmando que, debido a su excesivo énfasis en los protocolos y procedimientos, una revisión sistemática convencional degrada el estatus de las actividades académicas interpretativas, como leer, escribir y hablar, y las sustituye con una serie de tareas técnicas auditables. En opinión de esta autora, este cambio se debe en parte al nuevo gerencialismo en la investigación y da lugar a una «versión informatizada e impersonal de la síntesis de la investigación». Una vez escribí un breve comentario titulado «Why are the Cochrane Reviews so boring?» («¿Por qué las revisiones Cochrane son tan aburridas?»), argumentando que un enfoque excesivamente tecnocrático de la obtención y síntesis de los datos elimina el *significado* de una revisión[20]. Sin embargo, aunque esto puede ser cierto y MacLure puede tener parte de razón, hay que tratar de separar el grano de la paja. La revisión sistemática, bien empleada, salva vidas.

Bibliografía

1 Caveman A. The invited review? or, my field, from my standpoint, written by me using only my data and my ideas, and citing only my publications. Journal of Cell Science 2000;**113**(Pt 18):3125.

2 Pauling L. *How to live longer and feel better*. Portland, Oregon: Oregon State University Press; 1986 3125-3126.

3 McAlister FA, Clark HD, van Walraven C, et al. The medical review article revisited: has the science improved? Annals of InternalMedicine 1999;**131**(12):947-51.

4 Oxman AD, Guyatt GH. Thescience of reviewing research. Annals of theNewYork Academy of Sciences 1993;**703**(1):125-34.

5 Antman EM, Lau J, Kupelnick B, et al. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. JAMA:The Journal of the American Medical Association 1992;**268**(2):240-8.

6 Knipschild P, Systematic reviews. Some examples. BMJ: British Medical Journal 1994;**309**(6956):719-21.

7 Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of Internal Medicine 2009;**151**(4):264-9.

8 Bruins Slot KM, Berge E, Saxena R, et al. Oral anticoagulants versus antiplatelet therapy for preventing stroke and systemic embolic events in patients with atrial fibrillation. Cochrane Database of Systematic Reviews 2012;(Issue 2) Art. No.: CD009538. DOI: 10.1002/14651858.CD009538.

9 Greenhalgh T, Robert G, Macfarlane F, et al. Diffusion of innovations in service organizations: systematic review and recommendations. The Milbank Quarterly 2004;**82**(4):581-629 doi: 10.1111/j.0887-378X.2004.00325.x.

10 Morrison A, Polisena J, Husereau D, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. International Journal of Technology Assessment in Health Care 2012;**28**(2):138-44.

11 Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ: BritishMedical Journal 2005;**331**(7524):1064-5.

12 Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.1. 0 [updated March 2011]. Oxford: The Cochrane Collaboration 2011;.

13 Counsell CE, Clarke MJ, Slattery J, et al. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? BMJ: BritishMedical Journal 1994;**309**(6970):1677.

14 Cuijpers P, Hollon SD, van Straten A, et al. Does cognitive behaviour therapy have an enduring effect that is superior to keeping patients on continuation pharmacotherapy? A meta-analysis. BMJ Open 2013;**3**(4) doi: 10.1136/bmjopen-2012-002542[published Online First: Epub Date].

15 Egger M, Smith GD, Altman D. *Systematic reviews in health care: meta-analysis in context*. Chichester: Wiley.com; 2008.

16 Fergusson D, Glass KC, Hutton B, et al. Randomized controlled trials of aprotinin in cardiac surgery: could clinical equipoise have stopped the bleeding? Clinical Trials 2005;**2**(3):218-32.

17 Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. Evaluation & the Health Professions 2002;**25**(1):76-97.

18 Borenstein M, Hedges LV, Higgins JP, et al. *Introduction to meta-analysis*. Chichester: Wiley.com; 2011.
19 Thompson SG. Why and how sources of heterogeneity should be investigated. In: Egger M, Davey Smith G, Altman DG, editors. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Publications; 2001. p. 157-75.
20 Greenhalgh T. Outside the box: why are Cochrane reviews so boring? The British Journal of General Practice 2012;**62**(600):157-75 371.
21 Eysenck H. Problems withmeta-analysis. In: Chalmers I, Altman DG, editors. *Systematic reviews*. London: BMJ Publications; 1995.
22 Higgins JP, Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. International Journal of Epidemiology 2002;**31**(1):96-104.
23 Egger M, Smith GD. Misleading meta-analysis. BMJ: British Medical Journal 1995;**311**(7007):753-4.
24 Hennekens CH, DeMets D. The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. JAMA:The Journal of the American Medical Association 2009;**302**(21):2361-2.
25 Griffin S, Greenhalgh T. Diabetes care in general practice: meta-analysis of randomised control trials Commentary: meta-analysis is a blunt and potentially misleading instrument for analysing models of service delivery. BMJ: British Medical Journal 1998;**317**(7155):390-6.
26 Greenhalgh T, Commentary:. meta-analysis is a blunt and potentially misleading instrument for analysing models of service delivery. BMJ: British Medical Journal (Clinical research ed.) 1998;**317**(7155):395-6.
27 Devillé WL, Buntinx F, Bouter LM, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Medical Research Methodology 2002;**2**(1):9.
28 Wong G, Greenhalgh T, Westhorp G, et al. RAMESES publication standards: meta-narrative reviews. BMC Med 2013;11:20 doi: 10.1186/1741-7015-11-20.
29 Wong G, Greenhalgh T, Westhorp G, et al. RAMESES publication standards: realist syntheses. BMC Med 2013;11:20 doi: 10.1186/1741-7015-11-21.
30 MacLure M. 'Clarity bordering on stupidity': where's the quality in systematic review? Journal of Education Policy 2005;**20**(4):393-416.

Capítulo 10 Artículos que ofrecen pautas de actuación (guías)

El gran debate sobre las guías

Las actitudes respecto a las guías clínicas nunca han sido tan divergentes entre los médicos de la primera línea asistencial y quienes elaboran políticas desde los despachos. Los elaboradores de políticas (término en el que englobo a todos los que tienen una visión de cómo debería ejercerse la medicina en un mundo ideal, incluidos los políticos, directivos, directores clínicos, académicos y profesores) tienden a amar las guías. Los médicos de primera línea (es decir, las personas que dedican todo su tiempo a atender a los pacientes) a menudo sienten una fuerte aversión hacia estas guías.

Antes de seguir ahondando en este espinoso tema político, se necesita definir lo que son las guías, algo que puede hacerse del siguiente modo.

> *Las guías son afirmaciones elaboradas sistemáticamente para ayudar al médico a tomar decisiones sobre la asistencia sanitaria adecuada en circunstancias clínicas específicas.*

Una de mis colegas, la Dra. Deborah Swinglehurst, ha publicado recientemente un espléndido artículo sobre guías basadas en la evidencia (qué son, cómo se elaboran, por qué se necesitan y qué controversias existen al respecto)[1]. Me he basado en gran medida en su revisión a la hora de actualizar este capítulo. Deborah establece en su artículo una distinción importante entre las guías (que suelen expresarse en términos de principios generales y dejan un amplio margen para basarse en el criterio propio) y los protocolos, que define del siguiente modo: «Los protocolos son instrucciones sobre lo que se debe hacer en determinadas circunstancias. Son similares a las guías, pero dejan menos margen para el criterio individual, suelen elaborarse para personal menos experimentado o para su uso en situaciones donde las eventualidades son predecibles».

Los propósitos de las guías se presentan en el cuadro 10.1. Las reticencias de los clínicos hacia las guías se deben a varias razones[2,7]:

- Libertad clínica («no quiero que nadie me diga cómo tratar a mis pacientes»).
- Debates entre los expertos sobre la calidad de la evidencia («bueno, si no son capaces de ponerse de acuerdo entre ellos…»).

Cuadro 10.1 Finalidad de las guías

1. Hacer que las normas basadas en la evidencia sean explícitas y accesibles (pero v. el texto subsiguiente: pocas guías de las que se usan actualmente están verdaderamente basadas en la evidencia).
2. Lograr que la toma de decisiones en la clínica y a la cabecera del paciente sea más fácil y objetiva.
3. Proporcionar un punto de referencia para evaluar la actividad profesional.
4. Delimitar la división del trabajo (p. ej., entre los médicos generales y los especialistas).
5. Educar a los pacientes y a los profesionales sobre las mejores prácticas actuales.
6. Mejorar la rentabilidad de los servicios sanitarios.
7. Servir como herramienta para el control externo.

- Menosprecio de la evidencia por los médicos («todo esto está muy bien, pero, cuando yo estudiaba, siempre nos enseñaban a no administrar esteroides para el asma»).
- Medicina defensiva («pediré todas las pruebas de todos modos; hombre precavido vale por dos»).
- Limitaciones estratégicas y de costes («no podemos permitirnos reemplazar el equipo»).
- Limitaciones prácticas específicas («¿dónde he puesto esas guías?»).
- Reticencia de los pacientes a aceptar procedimientos («la paciente insiste en que necesita un frotis cervical cada año»).
- Entrada en conflicto con otros factores no médicos («cuando tengamos el nuevo sistema informático en funcionamiento…»).
- Falta de retroalimentación adecuada específica del paciente sobre el rendimiento («creo que estoy tratando adecuadamente esta enfermedad»).
- Confusión («la guía no parece que me ayude con el problema que estoy tratando»).

La imagen del bufón médico dando palos de ciego alegremente en las consultas externas diagnosticando las mismas enfermedades y prescribiendo los mismos fármacos que se aprendieron en la facultad de medicina hace 40 años y que nunca ha leído un artículo echa por tierra el argumento de la libertad clínica. Este tipo de situaciones hipotéticas son la fuerza motriz para quienes pretenden imponer las «guías de expertos» en la mayor parte o en toda la práctica médica, y sirven de ejemplo de todos aquellos que no se mantienen actualizados.

Sin embargo, existe un argumento poderoso que suele aducirse en contra del uso excesivo, y en particular de la imposición compulsiva, de las guías clínicas, que fue expresado elocuentemente por el profesor J. Grimley Evans hace algunos años[8]:

Existe el temor de que, en ausencia de evidencia claramente aplicable al caso que esté tratando, un clínico podría verse obligado por las guías a emplear una evidencia que sólo tenga una relevancia dudosa y que tal vez proceda de un grupo diferente de pacientes de otro país, de otro momento y en el que se usó un tratamiento similar pero no idéntico. Esto es medicina sesgada por la evidencia; consiste en utilizar la evidencia del modo en el que el borracho del chiste que buscaba las llaves de su casa bajo una farola porque es donde había luz, a pesar de que se le habían caído en otro lugar.

El temor de Grimley Evans, que todos los médicos comparten pero que pocos expresan, es que los políticos y los gestores de servicios sanitarios que se han subido al carro de la medicina basada en la evidencia (MBE) utilicen las guías para imponer el tratamiento de enfermedades en lugar que el de los pacientes. Se teme que sus juicios sobre las personas y sus enfermedades estén subordinados a la evidencia publicada de que una intervención es eficaz «en promedio». Este y otros inconvenientes reales y subjetivos de las guías se presentan en el cuadro 10.2, que se ha elaborado a partir de varias fuentes[2-6]. Sin embargo, si se ha leído la distinción indicada previamente entre guías y protocolos, se habrá observado que una buena guía no *obligaría* a abandonar el sentido común o el criterio personal sino que simplemente señalaría una pauta de actuación recomendada que se debería tener en cuenta.

Cuadro 10.2 Inconvenientes de las guías (reales y subjetivos)

1. Las guías pueden ser intelectualmente sospechosas y reflejar la «opinión de expertos», lo que puede institucionalizar una práctica poco sólida.
2. Al reducir la variación de la práctica médica, puedan estandarizar una «práctica promedio», en lugar de la mejor práctica.
3. Pueden inhibir la innovación y evitar que los casos individuales sean tratados de forma específica y adecuada.
4. Las guías desarrolladas a nivel nacional o regional pueden no reflejar las necesidades locales o no ser asumidas como propias por los profesionales locales.
5. Las guías elaboradas en atención secundaria pueden no reflejar las diferencias demográficas, clínicas o prácticas entre ese contexto y el de la atención primaria.
6. Las guías pueden producir modificaciones no deseables en el equilibrio de poder entre los diferentes grupos profesionales (p. ej., entre los médicos y los académicos o los clientes y proveedores). Por lo tanto, la elaboración de guías puede ser percibida como un acto político.
7. Las guías desactualizadas podrían evitar la implementación de la nueva evidencia procedente de la investigación.

Sin embargo, incluso una guía perfecta puede suponer una carga para un médico atareado. Mi amigo Neal Maskrey me envió recientemente esta cita de un artículo publicado en la revista *The Lancet:*

> *Se analizó una guardia (de 24 horas) de urgencias de medicina en nuestro hospital. En una guardia relativamente tranquila, se vio a 18 pacientes con un total de 44 diagnósticos. Las guías que el médico de guardia debería haber leído, recordado y aplicado correctamente para esos trastornos abarcaban 3.679 páginas. Esta cifra sólo incluía las guías NICE (National Institute for Health and Care Excellence de Reino Unido), las de los Royal Colleges y de las principales sociedades científicas de los últimos 3 años. Si se tardase 2 minutos en leer cada página, el médico de guardia debería haber dedicado 122 horas para mantenerse al tanto de las guías*[9].

La industria floreciente de la guías debe su éxito, al menos en parte, a una cultura creciente de «rendición de cuentas» que, según muchos autores alegan, se está constituyendo en la norma en muchos países. En el National Health Service (Servicio Nacional de Salud) de Reino Unido, todos los médicos, enfermeras, farmacéuticos y otros profesionales sanitarios tienen actualmente la obligación contractual de proporcionar asistencia clínica basada en la mejor evidencia disponible. Las guías elaboradas o validadas oficialmente, como las publicadas por el National Institute of Health and Care Excellence de Reino Unido (www.nice.org.uk) son una forma de respaldar y de controlar ese loable objetivo.

Mientras que las implicaciones médico-legales de las guías «oficiales» pocas veces se han evaluado en Reino Unido, los tribunales de Estados Unidos han dictaminado que los elaboradores de guías pueden ser declarados responsables de guías erróneas. Aún más preocupante es que un tribunal estadounidense se negó recientemente a aceptar que se había seguido una guía basada en la evidencia (que recomendaba a los médicos que compartiesen la incertidumbre inherente asociada con la determinación del antígeno prostático específico [PSA] en varones de mediana edad asintomáticos y que tomasen una decisión compartida sobre si merecía la pena realizar el análisis) como defensa para un médico que había sido demandado por no diagnosticar un cáncer de próstata precoz en un desafortunado paciente de 53 años[10].

¿Cómo se puede ayudar a garantizar que se siguen las guías basadas en la evidencia?

Dos de las principales autoridades internacionales en el espinoso tema de la aplicación de las guías clínicas son Richard Grol y Jeremy Grimshaw. En un estudio inicial del equipo de Grol, los principales factores asociados con el seguimiento satisfactorio de una guía o protocolo fueron la percepción de los médicos de que era indiscutible (68% de cumplimiento frente al 35% si se percibe como controvertida), basada en la evidencia (71% frente al 57% en caso contrario), si

contenía recomendaciones explícitas (67% frente al 36% si las recomendaciones eran vagas) y si no requería un cambio de las rutinas existentes (67% frente al 44% si se recomendaba un cambio importante)[7].

Otro artículo inicial, de Grimshaw y Russell[11], que se resume en la tabla 10.1, demostró que la probabilidad de que una guía se siga en la práctica dependía de tres factores:

(a) La estrategia de elaboración (dónde y cómo se elaboró la guía).
(b) La estrategia de difusión (cómo se presentó a los médicos).
(c) La estrategia de implementación (cómo se animó y se apoyó al clínico para seguir la guía, incluidos los aspectos organizativos).

En lo que respecta a la estrategia de elaboración, como se muestra en la tabla 10.1, las guías más eficaces se elaboran a nivel local por las personas que van a utilizarlas, se introducen como parte de una intervención educativa específica y se implementan a través de un indicador específico del paciente que aparece en el momento de la consulta. La importancia del sentido de propiedad (es decir, la sensación que tienen las personas a las que se pide que jueguen con nuevas reglas de haber participado en la elaboración de dichas reglas) es sin duda evidente. También hay una extensa literatura sobre la teoría de la gestión que respalda la idea de sentido común de que los profesionales se opondrán a los cambios que perciban como una amenaza a su sustento (es decir, ingresos), autoestima, sentido de la competencia o autonomía. Por lo tanto, es lógico que la participación de los profesionales sanitarios a la hora de establecer las normas con las que serán evaluados suele producir mayores cambios en los resultados de los pacientes de los que tendrían lugar si no estuvieran implicados.

Las conclusiones de Grimshaw derivadas de este primer trabajo fueron mal interpretadas inicialmente por algunas personas en el sentido de que no había

Tabla 10.1 Clasificación de las guías clínicas en términos de probabilidad de ser eficaces

Probabilidad de ser eficaz	Estrategia de elaboración	Estrategia de difusión	Estrategia de implementación
Alta	Interna	Intervención educativa específica (p. ej., programa de aprendizaje basado en el problema)	Recordatorio específico del paciente en el momento de la consulta
Por encima de la media	Intermedia	Formación continuada (p. ej., conferencia)	Retroalimentación específica del paciente
Por debajo de la media	Externa, local	Grupos de correo dirigidos	Retroalimentación general
Baja	Externa, nacional	Publicación en revistas	Recordatorio general

Fuente: Grimshaw y Rusell[11]. Reproducida con autorización de Elsevier.

lugar para las guías desarrolladas a nivel nacional porque sólo las elaboradas localmente tendrían algún impacto. De hecho, aunque la adopción local y el sentido de propiedad son, sin duda, cruciales para el éxito de un programa de guías, los equipos locales producen guías más sólidas si se basan en el conjunto de recursos nacionales e internacionales de recomendaciones basadas en la evidencia y lo utilizan como su punto de partida[12].

No se trata de que las aportaciones de los equipos locales reinventen la rueda en términos de resumir la evidencia, sino que tengan en cuenta aspectos prácticos locales a la hora de implementar la guía[12]. Por ejemplo, una guía elaborada a nivel nacional sobre el tratamiento de la epilepsia podría recomendar la presencia de una enfermera especialista en epilepsia en todas las áreas sanitarias. Sin embargo, es posible que en una de las áreas los equipos de atención sanitaria hubiesen ofertado una plaza de este tipo de enfermera, pero no hubiesen podido cubrirla, de modo que la «aportación local» podría consistir en determinar la mejor manera de proporcionar lo que la enfermera especialista en epilepsia habría aportado, en ausencia de una persona que cubriese dicho puesto.

En cuanto a la difusión y aplicación de las guías, el equipo de Grimshaw[13] publicó una revisión sistemática exhaustiva sobre las estrategias destinadas a mejorar la implementación de las guías por parte de los médicos en 2005.

Los resultados confirmaron el principio general de que los médicos no son fácilmente influenciables, pero que los esfuerzos para aumentar el uso de las guías suelen ser eficaces en cierto grado. De forma específica:

- Se demostraron mejoras en la dirección prevista de la intervención en el 86% de las comparaciones, pero el efecto fue generalmente de pequeña magnitud.
- Los recordatorios simples fueron la intervención en la que se observó una eficacia más constante.
- Los programas de divulgación educativa (p. ej., visitar a los médicos en sus clínicas) sólo tuvieron efectos modestos sobre el éxito de la implementación y resultaron muy caros en comparación con los enfoques menos intensivos.
- La difusión de materiales educativos produjo efectos modestos, pero potencialmente importantes (y de magnitud similar a intervenciones más intensivas).
- Las intervenciones multidisciplinarias no eran necesariamente más eficaces que las intervenciones individuales.
- No se podía extraer ninguna conclusión de la mayoría de los estudios primarios sobre la rentabilidad de la intervención.

La revisión de Grimshaw y cols. de 2005 revirtió algunas «ideas preconcebidas» previas, que probablemente eran el resultado de un sesgo de publicación en los ensayos de las estrategias de implementación. Al contrario de lo que yo afirmé en la primera y segunda ediciones de este libro, por ejemplo, las intervenciones complejas y caras destinadas a mejorar la implementación de las guías por parte de los médicos no suelen ser más eficaces que las guías sencillas, más baratas y bien orientadas. Se consideró que sólo el 27% de los estudios de intervención revisados por el equipo de Grimshaw estaban basados (implícita o explícitamente) en una teoría explícita del cambio. Dicho de otro modo, los investigadores de esos

estudios por lo general no basaron el diseño de su intervención en un mecanismo de acción adecuadamente articulado («A tiene como fin conseguir B que, a su vez, tiene como fin lograr C»).

En un artículo diferente, el equipo de Grimshaw[14] afirmó tajantemente que la investigación sobre la implementación de las guías debería basarse más en la teoría. Esa recomendación inspiró una corriente investigadora importante, que se ha resumido en un artículo de revisión de Eccles y cols.[15] y en una revisión sistemática de Davies y cols. sobre las estrategias de elaboración de guías basadas en la teoría[14]. En resumen, la aplicación de teorías sobre el cambio del comportamiento parece mejorar la asimilación de las guías por parte de los médicos, pero no es una garantía de éxito, por todas las razones que se expondrán en el capítulo 15[16].

Una de las contribuciones principales de Grimshaw a la MBE fue la creación de un subgrupo especial de la Cochrane Collaboration para revisar y resumir la investigación emergente sobre el uso de guías y otros asuntos relacionados con la mejora de la práctica profesional. Se pueden consultar detalles sobre el grupo Effective Practice and Organisation of Care (práctica y organización sanitaria eficaces, EPOC por su acrónimo en inglés) en la página web de Cochrane (http://www.epoc.cochrane.org/). En la base de datos del grupo EPOC se recogen miles de estudios primarios y más de 75 revisiones sistemáticas sobre el tema general de aplicar la evidencia de la investigación en la práctica.

Diez preguntas que deben plantearse sobre una guía clínica

Swinglehurst[1] señala acertadamente que todo lo que se ha proclamado a bombo y platillo para alentar a los médicos a seguir las guías sólo está justificado si la guía merece la pena seguirse como primera elección. Por desgracia, esto no es así para todas ellas. Esta autora sugiere dos aspectos de una buena guía: el contenido (p. ej., si se basa en una revisión sistemática exhaustiva y rigurosa de la evidencia) y el proceso (cómo se elaboró la guía). Personalmente añadiría un tercer aspecto: la presentación de la guía (lo atractiva que resulta al médico ocupado y lo fácil que es seguirla).

Al igual que todos los artículos publicados, las guías serían más fáciles de evaluar en todos estos aspectos si se presentaran en un formato estandarizado. Recientemente se ha publicado un estándar internacional (el instrumento AGREE, acrónimo de Appraisal of Guidelines for Research and Evaluation o Instrumento para la investigación y evaluación de guías de práctica clínica) para el desarrollo, publicación y presentación de las guías[17]. En el cuadro 10.3 se presenta una lista de comprobación pragmática, basada en parte en el trabajo del grupo AGREE, para la estructuración de la evaluación de una guía clínica; en el cuadro 10.4 se reproducen de forma completa los criterios AGREE revisados. Dado que pocas guías publicadas siguen actualmente este formato, es probable que deba escrutarse todo el texto para contestar las respuestas a las preguntas que se presentan aquí.

Cuadro 10.3 Marco esquemático para evaluar una guía clínica (v. también apéndice 1)

- *Objetivo:* el objetivo principal de la guía, incluidos el problema de salud y los pacientes, proveedores y contextos a los que se dirige.
- *Opciones:* opciones de práctica clínica tenidas en cuenta a la hora de elaborar la guía.
- *Resultados:* resultados de salud y económicos importantes tenidos en cuenta al comparar las prácticas alternativas.
- *Evidencia:* cómo y cuándo se recopiló, seleccionó y sintetizó la evidencia.
- *Valores:* descripción de cómo se asignaron valores a los posibles resultados de las opciones prácticas y quién participó en el proceso.
- *Beneficios, perjuicios y costes:* tipo y magnitud de beneficios, perjuicios y costes esperados para los pacientes al implementar las guías.
- *Recomendaciones:* resumen de las recomendaciones clave.
- *Validación:* descripción de cualquier revisión externa, comparación con otras guías o evaluación clínica del uso de la guía.
- *Patrocinadores y partes implicadas:* divulgación de las personas que desarrollaron, financiaron y respaldaron la guía.

Cuadro 10.4 Los seis dominios del instrumento AGREE II (v. referencia 16)

Dominio 1. Alcance y objetivo

1. El(los) objetivo(s) general(es) de la guía está(n) específicamente descrito(s).
2. La(s) pregunta(s) de salud cubierta(s) por la guía está(n) específicamente descrita(s).
3. La población a la cual se pretende aplicar la guía está específicamente descrita.

Dominio 2. Participación de los implicados

1. El grupo que desarrolla la guía incluye individuos de todos los grupos profesionales relevantes.
2. Se han tenido en cuenta los puntos de vista de la población diana.
3. Los usuarios diana de la guía están claramente definidos.

Dominio 3. Rigor en la elaboración

1. Se han utilizado métodos sistemáticos para la búsqueda de la evidencia.
2. Los criterios para seleccionar la evidencia se describen con claridad.
3. Las fortalezas y limitaciones del conjunto de la evidencia están claramente descritas.
4. Los métodos utilizados para formular las recomendaciones están claramente descritos.

5. Al formular las recomendaciones se han considerado los beneficios para la salud, los efectos secundarios y los riesgos.
6. Hay una relación explícita entre cada una de las recomendaciones y las evidencias en las que se basan.
7. La guía ha sido revisada por expertos externos antes de su publicación.
8. Se incluye un procedimiento para actualizar la guía.

Dominio 4. Claridad y presentación

1. Las recomendaciones son específicas y no son ambiguas.
2. Las distintas opciones para el tratamiento de la enfermedad o problema de salud se presentan claramente.
3. Las recomendaciones clave son fácilmente identificables.

Dominio 5. Aplicabilidad

1. La guía ofrece consejo o herramientas para apoyar su implementación.
2. La guía describe factores facilitadores y barreras para su adopción.
3. Se han considerado los costes potenciales de la aplicación de las recomendaciones.
4. La guía ofrece criterios para realizar la monitorización o auditoría.

Dominio 6. Independencia editorial

1. Los puntos de vista de la entidad financiadora no han influido en el contenido de la guía.
2. Se han registrado y abordado los conflictos de intereses de los miembros del grupo de desarrollo de la guía.

A la hora de preparar esta lista, me he basado en muchos de los artículos que aparecen en la bibliografía de este capítulo, así como en el instrumento AGREE relativamente nuevo.

Pregunta uno: ¿La preparación y publicación de esta guía conllevan un conflicto de intereses significativo?

No entraré en detalles, pero una compañía farmacéutica que fabrica un tratamiento hormonal sustitutivo o un profesor de investigación cuyo trabajo de toda su vida ha consistido en perfeccionar este tratamiento podría tener la tentación de recomendarlo para indicaciones más amplias que un médico promedio. Mucho se ha escrito acerca de la «medicalización» de la experiencia humana (¿son «hiperactivos» los niños enérgicos con un intervalo de atención corto?; ¿debería ofrecerse «tratamiento» a las mujeres con deseo sexual bajo?, etc.). Una guía puede estar basada en la evidencia, pero el problema al que se refiere habrá sido establecido por un equipo que contempla el mundo de una manera particular.

Pregunta dos: ¿La guía está relacionada con un tema apropiado e indica claramente el grupo diana al que se aplica?

En el cuadro 10.5 se presentan las preguntas clave relativas a la elección del tema, tomadas de un artículo publicado hace unos años en el *British Medical Journal*[18].

Cuadro 10.5 Preguntas clave sobre la elección del tema para la elaboración de la guía (v. referencia 17)

- ¿El tema se asocia a un alto volumen de pacientes, un alto riesgo y un alto coste?
- ¿Existen variaciones amplias o inexplicables en la práctica?
- ¿Es importante el tema en términos del proceso y el resultado de la asistencia al paciente?
- ¿Hay posibilidades de mejora?
- ¿Es probable que se vaya a reembolsar la inversión de tiempo y dinero?
- ¿Es probable que el tema suscite el interés de los miembros del equipo?
- ¿Es probable alcanzar un consenso?
- ¿El cambio beneficiará a los pacientes?
- ¿Se puede implementar el cambio?

La cita de Grimley Evans que se recoge en la página 137 suscita la pregunta: ¿a quién se aplica esta guía? Si la evidencia corresponde a personas de entre 18 y 65 años sin enfermedades concurrentes (es decir, sin ningún otro trastorno, excepto la enfermedad que se está considerando), podría no ser aplicable a nuestros pacientes. A veces, esto significa que deberemos rechazarla de plano, pero lo más frecuente es que haya que utilizar el criterio propio para evaluar su transferibilidad.

Pregunta tres: ¿El panel de desarrollo de la guía incluyó (a) un experto en el área temática, (b) un especialista en los métodos de investigación secundaria (p. ej., metaanalista, economista sanitario) y (c) una persona afectada por la enfermedad?

De forma paradójica, si una guía clínica la ha preparado en su totalidad un panel de «expertos» internos, debe contemplarse de un modo especialmente crítico, ya que los investigadores han demostrado ser menos objetivos a la hora de valorar la evidencia sobre su propio campo de experiencia que sobre el de otra persona. La participación de alguien ajeno (un experto en desarrollo de guías en lugar de en el tema clínico particular) que actúe como árbitro y asesor metodológico debería hacer que el proceso fuese más objetivo. Sin embargo, como Gabbay y cols.[19] demostraron en un estudio cualitativo, la experiencia difícil de evaluar (lo que podría denominarse *conocimiento incorporado*) de los médicos de primera línea (en este caso, los médicos de atención primaria) contribuyó decisivamente al desarrollo de guías locales manejables. Sin embargo, toda la experiencia objetiva del mundo no puede sustituir el hecho de ser uno mismo quien padece la enfermedad en cuestión y la evidencia emergente sugiere que los pacientes y los cuidadores aportan una tercera perspectiva crucial para el proceso de desarrollo de guías[20].

Pregunta cuatro: ¿Se han explicitado los criterios subjetivos del panel de desarrollo y están justificados?

El desarrollo de guías no es sólo un proceso técnico para encontrar evidencia, evaluarla y convertirla en recomendaciones. Las recomendaciones también requieren una serie de criterios (relacionados con los valores personales o sociales, principios éticos, etc.). Como el National Institute for Health and Care Excellence (NICE) británico ha afirmado (v. www.nice.org.uk), es correcto y apropiado que quienes elaboran guías tengan en cuenta los «principios éticos, preferencias, cultura y aspiraciones que deberían sustentar la naturaleza y el alcance de la asistencia prestada por el National Health Service». Swinglehurst[1] sugiere cuatro subpreguntas que se deben plantear sobre estos criterios subjetivos:

- ¿Qué *principios rectores* se han utilizado para decidir el grado de eficacia que debe tener una intervención (en comparación con sus perjuicios potenciales) antes de considerar su recomendación?
- ¿Qué *valores* han apuntalado las decisiones del panel sobre qué desarrollos de guías se deben priorizar?
- ¿Cuál es el *marco ético* con el que están trabajando los desarrolladores de guías, en particular, respecto a las cuestiones de justicia distributiva («racionamiento»)?
- Cuando han surgido desacuerdos entre los desarrolladores de guías, ¿qué *procesos explícitos* se han utilizado para resolver tales desacuerdos?

Pregunta cinco: ¿Se han analizado y evaluado rigurosamente todos los datos relevantes?

La validez académica de las guías depende (entre otras cosas) de si están respaldadas por estudios de investigación primarios de alta calidad y de la fuerza de la evidencia de esos estudios. En el nivel más básico, hay que valorar si se realizó algún análisis de la literatura o si las guías son simplemente una declaración de un panel selecto de expertos sobre la práctica preferida (es decir, guías de consenso). Si se ha analizado la literatura, hay que determinar si se realizó una búsqueda sistemática y, en tal caso, si se siguió en términos generales el método descrito en la sección «Evaluación de las revisiones sistemáticas». ¿Se incluyeron todos los artículos encontrados en la búsqueda o se utilizó un sistema de puntuación explícito (como GRADE[21]) para rechazar los de mala calidad metodológica y otorgar el peso extra que merecían a los de alta calidad? Lo ideal sería utilizar revisiones sistemáticas actualizadas como materia prima para la elaboración de guías. Sin embargo, en muchos casos, una búsqueda de la investigación rigurosa y relevante en la que basar las guías clínicas resulta infructuosa y los autores recurren inevitablemente a la «mejor evidencia» disponible o a la opinión de expertos.

Pregunta seis: ¿Se ha sintetizado correctamente la evidencia y están las conclusiones de la guía en consonancia con los datos en los que se basan?

Otro factor determinante clave de la validez de una guía es cómo se han combinado los diferentes estudios que contribuyen a ella (es decir, sintetizado) en el

contexto de las necesidades clínicas y políticas que se están abordando. Por un lado, una revisión sistemática y un metaanálisis podrían haber sido apropiados y, en el caso de haber realizado este último, se deberían haber abordado los aspectos de la probabilidad y la confianza de manera aceptable (v. sección «Resumen», cap. 4).

Sin embargo, no hay (y nunca habrá) revisiones sistemáticas suficientes que contemplen todas las eventualidades en la toma de decisiones clínicas y la elaboración de políticas. En muchas áreas, sobre todo las complejas, la opinión de expertos sigue siendo la mejor «evidencia» disponible y en tales casos los desarrolladores de guías deberían adoptar métodos rigurosos para garantizar que no es únicamente la voz del experto que habla más tiempo en las reuniones la que queda reflejada en las recomendaciones. Los grupos formales de desarrollo de guías suelen tener un conjunto explícito de métodos (v., por ejemplo, este documento del NICE británico[22]).

En un análisis reciente de tres guías «basadas en la evidencia» para la apnea obstructiva del sueño, se observó que en ellas se indicaban recomendaciones muy diferentes a pesar de estar basadas en un conjunto casi idéntico de estudios primarios. La principal razón de la discrepancia era que los expertos tendían a dar más peso a los estudios de su propio país[23].

Pregunta siete: ¿La guía tiene en cuenta las variaciones de la práctica médica y otras áreas controvertidas (p. ej., una atención óptima en respuesta a una falta de financiación real o subjetiva)?

Sería poco prudente realizar afirmaciones dogmáticas acerca de la práctica ideal sin tomar como referencia lo que verdaderamente sucede en el mundo real. Hay muchos casos en los que algunos médicos marchan a un paso completamente diferente al del resto de los profesionales sanitarios (v. sección «¿Por qué hay quien se queja cuando oye hablar de medicina basada en la evidencia?») y una buena guía debe encarar estas realidades de frente en vez de esperar a que la minoría equivocada adopte el paso de los demás espontáneamente.

Otro tema espinoso que las guías deben afrontar es dónde se deben establecer compromisos esenciales si las limitaciones económicas impiden una práctica «ideal». Si el ideal, por ejemplo, es ofrecer a todos los pacientes con una arteriopatía coronaria significativa una operación de revascularización miocárdica (en el momento de escribir este libro, esto no es así, pero no importa) y el servicio de salud sólo puede financiar el 20% de dichos procedimientos, ¿a qué pacientes debe otorgarse la prioridad?

Pregunta ocho: ¿Es la guía clínicamente relevante, exhaustiva y flexible?

Dicho de otro modo, ¿está escrita desde la perspectiva del médico, enfermera, matrona, fisioterapeuta, etc., y tiene en cuenta el tipo de pacientes que es probable que atienda el profesional, y en qué circunstancias? Tal vez la fuente más frecuente de problemas a este respecto surge cuando las guías que se han elaborado en atención secundaria dirigidas a su uso en pacientes ambulatorios de hospital (que tienden a presentar una enfermedad más grave del espectro clínico) se proporcionan al equipo de atención primaria con la

intención de que se utilicen en el ámbito de la atención primaria, donde, en general, los pacientes están menos enfermos y pueden necesitar menos pruebas y un tratamiento menos agresivo. Este tema se comenta en la sección «Validación de las pruebas diagnósticas frente a un patrón oro» en relación con las diferentes utilidades de las pruebas diagnósticas y de cribado en distintas poblaciones.

Las guías deben abarcar la totalidad, o la mayoría, de las eventualidades clínicas. ¿Qué sucede si el paciente es intolerante a la medicación recomendada? ¿Qué pasa si no se pueden realizar todos los análisis de sangre recomendados? ¿Qué ocurre si el paciente es muy joven, muy viejo o presenta una enfermedad coexistente? Al fin y al cabo, éstos son los pacientes que obligan a la mayoría de los médicos a utilizar las guías, mientras que los pacientes más típicos tienden a ser tratados sin tener que recurrir a instrucciones escritas. Un trabajo reciente del equipo de Shekelle[2] ha añadido un factor crucial (la multimorbilidad) a las dificultades para seguir las guías: en ocasiones, el paciente tiene una enfermedad que impide aplicar el tratamiento estándar recomendado. De ello se desprende que los desarrolladores de guías deberían tener siempre en cuenta las enfermedades concurrentes a la hora de establecer sus recomendaciones.

La flexibilidad es una consideración especialmente importante para los organismos nacionales y regionales que se quieran dedicar a elaborar guías. Como se señaló anteriormente, el sentido de propiedad de las guías por parte de las personas que se pretende que las usen a nivel local es crucial para que se usen realmente. Si los profesionales no tienen total libertad para adaptarlas a las necesidades y prioridades locales, es probable que las guías nunca se saquen del cajón.

Pregunta nueve: ¿La guía tiene en cuenta lo que es aceptable, asequible y posible en la práctica para los pacientes?

Hay una historia apócrifa de un médico de la década de 1940 (una época en la que no se disponía de fármacos eficaces para la hipertensión arterial), que descubrió que si se limitaba la dieta de los pacientes hipertensos a arroz blanco hervido sin sal se reducía drásticamente su presión arterial y también se disminuía el riesgo de accidente cerebrovascular. Sin embargo, la historia acaba confirmando que la dieta causaba tal depresión en los pacientes que muchos de ellos se suicidaron.

Éste es un ejemplo extremo, pero en los últimos años he visto guías para tratar el estreñimiento en los ancianos cuya única alternativa consiste en la desagradable combinación de ingerir grandes cantidades de salvado y administrar supositorios dos veces al día. No es de extrañar que las enfermeras a las que se proporcionó estas directrices (que merecen todos mis respetos) hayan vuelto a administrar aceite de ricino.

En una revisión reciente, se puede consultar una exposición más detallada sobre cómo incorporar las necesidades y prioridades de los pacientes en el desarrollo de guías[20].

Pregunta diez: ¿La guía incluye recomendaciones para su propia difusión, aplicación y revisión periódica?

Dada la brecha bien documentada entre lo que se sabe que es una buena práctica y lo que realmente sucede (v. texto precedente), y las dificultades para la implementación satisfactoria de las guías descritas en la sección «¿Cómo se puede ayudar a garantizar que se siguen las guías basadas en la evidencia?», sería beneficioso para quienes desarrollan guías que sugiriesen métodos para maximizar su uso. Si este objetivo se incluyese sistemáticamente en las «Guías para elaborar buenas guías», es probable que los escritores de guías incluyesen menos ideales inalcanzables en sus recomendaciones y más consejos plausibles, viables y posibles de explicar a los pacientes. Dicho esto, un avance muy positivo en MBE desde que se publicó la primera edición de este libro es el cambio en las actitudes de los desarrolladores de guías: ahora suelen asumir la responsabilidad de relacionar sus recomendaciones con los médicos (y pacientes) del mundo real y de revisar y actualizar dichas recomendaciones periódicamente.

Bibliografía

1 Swinglehurst D. Evidence-based guidelines: the theory and the practice. Evidence-Based Healthcare and Public Health 2005;**9**(4):308-14.
2 Shekelle P, Woolf S, Grimshaw JM, et al. Developing clinical practice guidelines: reviewing, reporting, and publishing guidelines; updating guidelines; and the emerging issues of enhancing guideline implementability and accounting for comorbid conditions in guideline development. Implementation Science 2012;**7**(1):62.
3 Gurses AP, Marsteller JA, Ozok AA, et al. Using an interdisciplinary approach to identify factors that affect clinicians' compliance with evidence-based guidelines. Critical Care Medicine 2010;**38**:S282-91.
4 Gagliardi AR, Brouwers MC, Palda VA, et al. Howcanwe improve guideline use?A conceptual framework of implementability. Implementation Science 2011;**6**(1):26.
5 Evans-Lacko S, Jarrett M, McCrone P, et al. Facilitators and barriers to implementing clinical care pathways. BMC Health Services Research 2010;**10**(1):182.
6 Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. BMJ: British Medical Journal 2004;**328**(7435):343.
7 Grol R, Dalhuijsen J, Thomas S, et al. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. BMJ: British Medical Journal 1998;**317**(7162):858-61.
8 Evans JG. Evidence-based and evidence-biased medicine. Age and Ageing 1995;**24**(6):461-3.
9 Allen D, Harkins K. Too much guidance? The Lancet 2005;**365**(9473):1768.
10 Merenstein D. Winners and losers. JAMA: The Journal of the American Medical Association 2004;**291**(1):15-6.
11 Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. The Lancet 1993;**342**(8883):1317-22.
12 Harrison MB, Légaré F, Graham ID, et al. Adapting clinical practice guidelines to local context and assessing barriers to their use. Canadian Medical Association Journal 2010;**182**(2):E78-84.

13 Grimshaw J, Thomas R, MacLennan G, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. International Journal of Technology Assessment in Health Care 2005;**21**(01):149.
14 Eccles M, Grimshaw J, Walker A, et al. Changing the behavior of healthcare professionals: the use of theory in promoting the uptake of research findings. Journal of Clinical Epidemiology 2005;**58**(2):107-12.
15 EcclesMP, Grimshaw JM, MacLennan G, et al. Explaining clinical behaviors using multiple theoretical models. Implementation Science 2012;**7**:99.
16 Davies P, Walker AE, Grimshaw JM. A systematic review of the use of theory in the design of guideline dissemination and implementation strategies and interpretation of the results of rigorous evaluations. Implementation Science 2010;**5**:14.
17 BrouwersMC, Kho ME, Browman GP, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. Canadian Medical Association Journal 2010;**182**(18):E839-42.
18 Thomson R, Lavender M, Madhok R. How to ensure that guidelines are effective. BMJ: British Medical Journal 1995;**311**(6999):237-42.
19 Gabbay J, May Al. Evidence based guidelines or collectively constructed "mindlines?" Ethnographic study of knowledge management in primary care. BMJ: British Medical Journal 2004;**329**(7473):1013.
20 Boivin A, Currie K, Fervers B, et al. Patient and public involvement in clinical guidelines: international experiences and future perspectives. Quality and Safety in Health Care 2010;**19**(5):1-4.
21 Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. Journal ofClinical Epidemiology 2011;**64**(4):383-94.
22 Hill J, Bullock I, Alderson P. A summary of themethods that theNationalClinical Guideline Centre uses to produce clinical guidelines for the National Institute for Health and Clinical Excellence. Annals of InternalMedicine 2011;**154**(11):752-7.
23 AartsMC, van der Heijden GJ, Rovers MM, et al. Remarkable differences between three evidence-based guidelines on management of obstructive sleep apneahypopnea syndrome. The Laryngoscope 2013;**123**(1):283-91.

Capítulo 11 Artículos que indican lo que cuestan las cosas (análisis económicos)

¿Qué es un análisis económico?

Un análisis económico se puede definir como *aquel que implica el uso de técnicas analíticas para definir las opciones a la hora de asignar los recursos.* La mayor parte de mi exposición sobre este tema procede de los consejos elaborados por el equipo del profesor Michael Drummond[1] para los autores y revisores de análisis económicos, así como del excelente resumen de bolsillo de Jefferson y cols.[2], quienes subrayan la importancia de enmarcar las cuestiones económicas sobre un artículo en el contexto de la calidad global y la pertinencia del estudio (v. sección «Diez preguntas que se deben plantear sobre un análisis económico»).

La primera evaluación económica que recuerdo fue un anuncio de televisión en el que el cantante de pop Cliff Richard trataba de convencer a un ama de casa de que la marca más cara de lavavajillas líquido del mercado «realmente resultaba más barata». Al parecer, era más enérgico contra las manchas, más suave para las manos y producía más espuma por cada céntimo que un «típico lavavajillas líquido barato». Aunque yo sólo tenía nueve años en aquella época, no acababa de estar convencida. ¿Con qué «típico lavavajillas líquido barato» se había comparado el producto? ¿Cuánto más enérgico era contra las manchas? ¿Por qué la eficacia de un lavavajillas líquido tenía que medirse en términos de espuma producida en lugar de hacerlo en platos limpiados?

Pido perdón a los lectores por insistir en este ejemplo trivial, pero me gustaría usarlo para ilustrar los cuatro tipos principales de evaluación económica que se encuentran en la literatura (v. las definiciones convencionales en la tabla 11.1):

- *Análisis de minimización del coste:* «la marca A cuesta 47 céntimos por envase, mientras que la marca B cuesta 63 céntimos por envase».
- *Análisis de coste-efectividad:* «la marca A lava 15 platos adicionales por lavado en comparación con la marca B».
- *Análisis de coste-utilidad:* «en términos de horas de ama de casa ajustadas por calidad (una puntuación compuesta que refleja el tiempo y el esfuerzo necesarios para dejar los platos limpios y la aspereza de las manos causada por

el lavavajillas líquido), la marca A produce 29 unidades por euro, mientras la marca B produce 23 unidades».

- *Análisis de coste-beneficio:* «El coste neto total (que refleja el coste directo del producto, el coste indirecto del tiempo dedicado a lavar los platos y el valor económico estimado de un plato limpio respecto a uno un poco sucio) del producto A por día es de 7,17 céntimos, mientras que el del producto B es de 9,32 céntimos».

El lector debería ser capaz de captar de inmediato que el análisis más apropiado que se debería utilizar en este ejemplo es el de coste-efectividad. El análisis de minimización del coste (v. tabla 11.1) es inadecuado, pues las marcas A y B no tienen la misma efectividad. El análisis de coste-utilidad es innecesario porque, en este ejemplo, nos interesan muy pocas cosas, aparte de la cantidad de platos limpiados por unidad de líquido lavavajillas. Dicho de otro modo, nuestro resultado sólo tiene una dimensión importante. En este ejemplo, el análisis de coste-beneficio es una forma absurdamente complicada de averiguar que la marca A limpia más platos por céntimo.

Sin embargo, hay muchas situaciones en las que los profesionales sanitarios, en particular los que proporcionan asistencia sanitaria a partir de presupuestos reales limitados, deben escoger entre intervenciones para una serie de trastornos diferentes cuyos resultados (como los casos de sarampión prevenidos, mayor movilidad tras una artroplastia total de cadera, menor riesgo de mortalidad por ataque cardíaco o probabilidad de dar a luz a un bebé vivo) no se pueden comparar directamente entre sí. Existe controversia no sólo acerca de cómo se deben realizar estas comparaciones (v. sección «Medición de los costes y beneficios de las intervenciones sanitarias»), sino también de quién debería realizarlas y a quién deberían rendir cuentas quienes toman las decisiones para el «racionamiento» de la asistencia sanitaria. Estas preguntas esenciales, fascinantes y frustrantes están más allá del ámbito de este libro, pero quien tenga interés puede consultar la obra reciente de Donaldson y Mitton[3].

Medición de los costes y beneficios de las intervenciones sanitarias

Hace unos años, me ingresaron en el hospital para una apendicectomía. Desde el punto de vista del hospital, el coste de mi asistencia incluía mi alojamiento y manutención durante 5 días, una parte del tiempo de los médicos y enfermeras, los fármacos y apósitos, así como las pruebas complementarias (análisis de sangre y TC). Otros *costes directos* (v. cuadro 11.1) fueron el tiempo de mi médico general para asistirme de urgencia por la noche y el coste de la gasolina que gastaba mi marido cuando me visitaba (por no hablar de los bombones y las flores).

A esto hay que añadir los *costes indirectos* de mi pérdida de productividad. Estuve de baja laboral durante 3 semanas y mis tareas domésticas se distribuyeron temporalmente entre varios amigos, vecinos y una joven niñera. Además, desde

Tabla 11.1 Tipos de análisis económico

Tipo de análisis	Medida de resultado	Condiciones de uso	Ejemplo
Análisis de minimización del coste	Sin medida de resultado	Se usa cuando se sabe (o se puede asumir) que el efecto de ambas intervenciones es idéntico	Comparar el precio de un fármaco de marca con su equivalente genérico si se ha demostrado su bioequivalencia
Análisis de coste-efectividad	Unidades naturales (p. ej., años de vida ganados)	Se usa cuando el efecto de las intervenciones puede expresarse en términos de una variable principal	Comparar dos tratamientos preventivos para una enfermedad por lo demás mortal
Análisis de coste-utilidad	Unidades de utilidad (p. ej., años de vida ajustados por calidad)	Se usa cuando el efecto de las intervenciones sobre el estado de salud tiene dos o más dimensiones importantes (p. ej., beneficios y efectos secundarios de los fármacos)	Comparar los beneficios de dos tratamientos para venas varicosas en términos de resultado quirúrgico, aspecto estético y riesgo de evento adverso grave (p. ej., embolia pulmonar)
Análisis de coste-beneficio	Unidades monetarias (p. ej., coste estimado de pérdida de productividad)	Se usa cuando es deseable comparar una intervención para esta enfermedad con una intervención para una enfermedad diferente	Para una autoridad de compras, decidir si financiar un programa de trasplante cardíaco o una planta de rehabilitación de ictus

Cuadro 11.1 Ejemplos de costes y beneficios de las intervenciones sanitarias

Costes	Beneficios
Directos	***Económicos***
Manutención y alojamiento	Prevención de enfermedades caras de tratar
Fármacos, apósitos, etc.	Evitación de ingresos hospitalarios
Pruebas complementarias	Reanudación del trabajo remunerado
Salarios del personal	
	Clínicos
Indirectos	Retraso del fallecimiento o la discapacidad
Días de trabajo perdidos	Alivio del dolor, náuseas, disnea, etc.
Valor del trabajo «no remunerado»	Mejora de la visión, audición, fuerza muscular, etc.
Intangibles	***Calidad de vida***
Dolor y sufrimiento	Aumento de la movilidad e independencia
Estigma social	Mejora del bienestar
	Liberación del papel de enfermo

mi punto de vista, había varios costes *intangibles,* como el malestar, la pérdida de la independencia, el exantema alérgico que desarrollé por la medicación y la antiestética cicatriz que ahora tengo en el abdomen.

Como se muestra en el cuadro 11.1, estos costes directos, indirectos e intangibles constituyen uno de los términos de la ecuación coste-beneficio. En el lado de los beneficios, la operación incrementó en gran medida mis probabilidades de seguir viva. Además, tuve un agradable descanso del trabajo y, siendo sincera, me gustó la atención y simpatía que recibí. (Se debe tener en cuenta que el «estigma social» de la apendicitis puede ser positivo. Sería menos probable que presumiese de mi experiencia si mi ingreso se hubiese debido, por ejemplo, a un ataque epiléptico o una crisis nerviosa, que tienen estigmas sociales negativos.)

En el ejemplo de la apendicitis, pocos pacientes percibirían una gran libertad de elección a la hora de decidir si optan por la operación. Sin embargo, la mayoría de las actuaciones sanitarias no implican procedimientos definitivos para enfermedades que amenazan la vida. La mayoría de las personas pueden contar con que desarrollarán al menos una enfermedad crónica, incapacitante y progresiva, como cardiopatía isquémica, hipertensión arterial, artritis, bronquitis crónica, cáncer, reumatismo, hipertrofia prostática o diabetes. En algún momento, casi todos nosotros nos veremos obligados a decidir si «resulta útil» someterse a una operación rutinaria, tomar un fármaco en particular o modificar nuestro estilo de vida (reducir nuestro consumo de alcohol o seguir una dieta para reducir el colesterol).

Está bien que las personas informadas tomen decisiones sobre su propia asistencia siguiendo una reacción visceral («prefiero vivir con mi hernia a que me operen» o «soy consciente del riesgo de trombosis, pero quiero seguir fumando y tomando la píldora [anticonceptiva]»). Sin embargo, cuando se trata de elegir sobre la asistencia de otras personas, los valores y prejuicios personales son lo último que debería formar parte de la ecuación. La mayoría de las personas quieren que los planificadores y los elaboradores de políticas utilicen criterios objetivos, explícitos y defendibles a la hora de tomar decisiones como: «no, esta paciente no puede recibir un trasplante renal».

Una forma importante de abordar la pregunta: «¿cuál es su utilidad?» para un estado de salud determinado (como la diabetes o el asma mal controlada) es preguntar cómo se siente a alguien que lo presente. Se han desarrollado varios cuestionarios que intentan medir el estado general de salud, como el Nottingham Health Profile, el cuestionario de salud general SF-36 (muy utilizado en Reino Unido) y el McMaster Health Utilities Index Questionnaire (popular en Estados Unidos). Se puede consultar un resumen de estos cuestionarios en este manual de referencia[4].

En algunas circunstancias, las medidas de bienestar específicas de la enfermedad son más válidas que las medidas generales. Por ejemplo, una respuesta afirmativa a la pregunta: «¿está muy preocupado por la comida que toma?» podría indicar ansiedad en una persona sin diabetes, pero una actitud normal de autocuidado en una persona con diabetes[5]. También ha habido un aumento del interés por las medidas de calidad de vida *específicas del paciente* para permitir que distintos pacientes asignen diferentes valores a aspectos particulares de su salud y bienestar. Cuando se analiza la calidad de vida desde el punto de vista del paciente, éste es un enfoque razonable y humano. Sin embargo, un economista sanitario tiende a tomar decisiones sobre grupos de pacientes o poblaciones, en cuyo caso las medidas de calidad de vida específicas del paciente, e incluso específicas de la enfermedad, tienen una relevancia limitada. Los lectores que quieran ponerse al día rápidamente sobre el debate activo acerca de la forma de medir la calidad de vida relacionada con la salud pueden consultar varias de las referencias que figuran al final de este capítulo[4,6-8].

Los autores de los instrumentos estándar (como el SF-36) para medir la calidad de vida a menudo han dedicado años a garantizar que sean válidos (es decir, que midan lo que creemos que están midiendo), fiables (que lo hagan siempre) y que respondan al cambio (es decir, si una intervención mejora o empeora la salud del paciente, la escala lo reflejará). Por esta razón, hay que sospechar seriamente de un artículo que evite esos instrumentos estándar y que use en su lugar una escala simple y elemental propia de los autores («la capacidad funcional se clasificó como buena, regular o mala según la impresión global del médico» o «se preguntó a los pacientes que puntuasen tanto su dolor como su nivel global de energía de uno a diez y se sumaron los resultados»). También se debe tener en cuenta que incluso los instrumentos que han sido aparentemente bien validados a menudo no se sostienen ante una evaluación rigurosa de su validez psicométrica[8].

Otra forma de abordar la pregunta: «¿cuál es su utilidad?» referida a estados de salud particulares es a través de los *valores de preferencia del estado de salud,* es decir, el valor que, en una situación hipotética, una persona sana otorgaría a un deterioro específico de su salud o que una persona enferma otorgaría a una recuperación de la salud[9]. Hay tres métodos principales para la asignación de estos valores:

- *Escala visual analógica:* se pide al participante que haga una marca en una línea fija, en uno de cuyos extremos está escrito, por ejemplo, «salud perfecta» y en el otro, «muerte», para indicar dónde situaría el estado en cuestión (p. ej., estar postrado en silla de ruedas por una artritis de cadera).
- *Negociación de tiempo:* se pide al participante que considere un estado de salud en particular (p. ej., infertilidad) y que estime cuántos de sus años restantes de plena salud sacrificaría para «curarse» de la enfermedad.
- *Apuesta estándar:* se pide al participante que considere la posibilidad de elegir entre vivir el resto de su vida en un estado de salud en particular o «hacer una apuesta» (p. ej., una operación) con una determinada probabilidad de éxito, que le devolvería al estado de salud plena si tuviese éxito, pero que causaría la muerte si fallase. A continuación, se modifican las probabilidades para ver en qué momento se decide que no merece la pena hacer la apuesta.

El año de vida ajustado por calidad (AVAC, QALY, *quality-adjusted life-year*) se puede calcular multiplicando el valor de preferencia para ese estado por el tiempo que es probable que el paciente pase en ese estado. Los resultados de los análisis de coste-beneficio se suelen expresar en términos de «coste por AVAC». Algunos ejemplos de ello se muestran en el cuadro 11.2[10-15] El coste absoluto por AVAC es a veces menos importante en la toma de decisiones que el grado de diferencia entre el coste por AVAC de un tratamiento antiguo barato y uno nuevo caro. Es posible que el nuevo fármaco sólo sea ligeramente más eficaz, pero mucho más caro. El valor que se utiliza para comparar si el beneficio «es útil» se denomina relación coste-efectividad incremental o RCEI (ICER en inglés). Un buen ejemplo de ello es la reciente introducción del dabigatrán (un anticoagulante caro, pero menos incómodo para el paciente que la warfarina, ya que requiere menos análisis de sangre), cuya RCEI en comparación con la warfarina se ha estimado en 13.957 £[16].

Hasta hace unos años, una de mis numerosas «tareas de comité» consistía en sentarme en el comité de evaluaciones del NICE (National Institute for Health and Care Excellence británico, que asesora al Department of Health sobre la relación coste-efectividad de los fármacos). Es casi imposible que los miembros de ese comité multidisciplinario entablen un debate sobre si se debe recomendar la financiación de un fármaco controvertido sin que surjan grandes diferencias de opinión y afloren las emociones. Por lo general, los datos de AVAC de alta calidad tienden a aclarar las cosas y apaciguar los ánimos en esos debates. Así pues, cualquier medida de valores de preferencia del estado de salud refleja las preferencias y prejuicios de las personas que han contribuido a su desarrollo. De hecho, es posible obtener valores diferentes de AVAC dependiendo de cómo se

Cuadro 11.2 Coste por AVAC (v. referencias 10-15)

Se debe tener en cuenta que estos son precios de 2013 en Reino Unido, por lo que los valores absolutos ya no son válidos, aunque sirven como valores relativos útiles para las condiciones de ejemplo.

Tratamiento con estatinas en la nefropatía crónica (en pacientes con riesgo cardiovascular basal alto)	1.073 £
Tratamiento con estatinas en la nefropatía crónica (en pacientes con riesgo cardiovascular basal bajo)	98.000 £
Traslado precoz a un centro especializado en neurociencia para los traumatismos craneoencefálicos	11.000 £
Apoyo para el cambio de estilo de vida en la diabetes tipo 2	6.736 £
Tratamiento del virus de la hepatitis C en adictos a drogas por vía parenteral	6.803 £
Mamoplastia de reducción en mujeres con mamas voluminosas y pesadas	1.054 £
Tratamiento sustitutivo de nicotina para dejar de fumar	973-2.918 £
Asesoramiento para dejar de fumar	440-1.319 £
Telesalud en las personas mayores con multimorbilidad	88.000 £

planteen las preguntas de las que derivan los valores de preferencia del estado de salud[17].

Como el especialista en ética médica John Harris ha señalado, los AVAC son, al igual que la sociedad que los produce, inherentemente discriminatorios en función de la edad, el sexo y la raza, y cargan contra las personas con discapacidades permanentes (porque incluso una curación completa de una enfermedad no relacionada no devolvería al individuo a una salud perfecta). Además, los AVAC distorsionan nuestros instintos éticos al centrar nuestras mentes en los años de vida en lugar de en la vida de las personas. Según alega Harris, a un lactante prematuro discapacitado que requiera una cuna de cuidados intensivos se le asignarán más recursos de los que merece en comparación con una mujer de 50 años con cáncer debido a que el lactante, si sobreviviese, tendría muchos más años de vida para ajustar por calidad[18].

Existen alternativas cada vez más confusas a los AVAC[4,6,19,20]. Algunas de las que estaban de moda cuando se terminó de escribir este libro son:

- Equivalentes a años saludables (EAS o HYE, *Healthy Years Equivalent*). Es una medida de tipo AVAC que incorpora la mejora o deterioro probable del estado de salud de la persona en el futuro.
- Disposición a pagar (DAP o WTP, *Willingness to Pay*) o disposición a aceptar (DAA, WTA, *Willingness to Accept*). Son medidas de la cantidad de personas que estarían dispuestas a pagar para obtener ciertos beneficios o evitar ciertos problemas.
- Años de vida ajustados por discapacidad (AVAD, o DALY, *Disability-Adjusted Life Year*), que se usan principalmente en los países en vías de desarrollo para

evaluar la carga global de las enfermedades crónicas y la privación. Es una medida cada vez más utilizada que no está exenta de críticas.

- TWiST (acrónimo en inglés de tiempo sin síntomas de la enfermedad y de toxicidad del tratamiento) y Q-TWiST (TWiST ajustado por calidad), quizá las más extrañas.

Mi consejo sobre todas estas medidas es que se revise cuidadosamente en qué consiste el número que se supone que es un indicador «objetivo» del estado de salud de una persona (o de una población) y la forma en que las diferentes medidas podrían variar de acuerdo con diferentes estados de enfermedad. En mi opinión, todas ellas tienen usos potenciales, pero ninguna es una medida absoluta o irrefutable de salud o enfermedad. (Hay que tener en cuenta también que yo no pretendo ser una experta sobre ninguna de estas medidas o sobre la forma de calcularlas, por lo que incluyo una amplia lista de referencias al final de este capítulo.)

Sin embargo, hay otra forma de análisis que, aunque no elimina la necesidad de utilizar valores numéricos arbitrarios sobre la vida y la integridad física, evita pasar la pelota al tejado de los economistas sanitarios. Este enfoque, denominado *análisis de coste-consecuencias,* presenta los resultados del análisis económico de una forma desglosada. Dicho de otro modo, expresa los diferentes resultados en función de sus distintas unidades naturales (es decir, algo real, como los meses de supervivencia, las extremidades amputadas o los bebés nacidos sanos), de modo que las personas puedan asignar sus propios valores a determinados estados de salud antes de comparar dos intervenciones bastante diferentes (p. ej., el tratamiento de la infertilidad frente a reducir el colesterol, como en el ejemplo que se mencionó en el cap. 1). El análisis de coste-consecuencias permite que los valores de preferencia del estado de salud tanto de los individuos como de la sociedad cambien con el tiempo y es especialmente útil cuando éstos estén cuestionados o es probable que cambien. Este enfoque también puede permitir que grupos o sociedades distintos de aquéllos en los que se realizó el ensayo original utilicen el análisis.

Diez preguntas que se deben plantear acerca de un análisis económico

La lista de comprobación elemental que se presenta a continuación se basa en gran medida en las fuentes mencionadas en el primer párrafo de este capítulo. Recomiendo encarecidamente que para obtener una lista más completa se consulten esas fuentes, sobre todo las recomendaciones oficiales del grupo de trabajo del BMJ[1].

Pregunta uno: ¿El análisis se basa en un estudio que responde a una pregunta clínica claramente definida sobre un tema de importancia económica?

Antes de tratar de asimilar lo que dice un artículo sobre costes, escalas de calidad de vida o utilidades, hay que asegurarse de que el ensayo que se está analizando es científicamente relevante y capaz de ofrecer respuestas no sesgadas y sin ambigüedades a la pregunta clínica planteada en su introducción

(v. cap. 4). Además, si hay pocas alternativas entre las intervenciones en términos de costes o beneficios, es probable que un análisis económico detallado sea inútil.

Pregunta dos: ¿Desde qué punto de vista se consideran los costes y beneficios?

Desde el punto de vista del paciente, él o ella suele querer mejorar lo antes posible. Desde el punto de vista de la hacienda pública, la intervención sanitaria más coste-eficaz es la que devuelve a todos los ciudadanos sin demora a la condición de contribuyentes y, cuando este estado ya no es sostenible, provoca su muerte súbita inmediata. Desde el punto de vista de la compañía farmacéutica, sería difícil imaginar una ecuación de coste-beneficio que no incluyese uno de los productos de la compañía y, desde el punto de vista de un fisioterapeuta, la eliminación de un servicio de fisioterapia nunca sería coste-eficaz. Ningún análisis económico carece de perspectiva. La mayoría asume la perspectiva del propio sistema de salud, aunque algunos tienen en cuenta los costes ocultos para el paciente y la sociedad (p. ej., como resultado de días de trabajo perdidos). No hay una perspectiva «correcta» para una evaluación económica, pero el artículo debe indicar claramente a quién corresponden los costes y beneficios que se han tenido en cuenta o se han descartado.

Pregunta tres: ¿Se ha demostrado que las intervenciones que se comparan son clínicamente eficaces?

Nadie quiere un tratamiento barato si no funciona. El artículo que se está leyendo puede ser simplemente un análisis económico, en cuyo caso se basará en un ensayo clínico publicado previamente o será una evaluación económica de un nuevo ensayo, cuyos resultados clínicos se presentan en el mismo artículo. En cualquier caso, habrá que asegurarse de que la intervención más barata no es sustancialmente menos eficaz en términos clínicos que la que va a rechazarse por razones de coste. (Sin embargo, hay que tener en cuenta que en un sistema sanitario con recursos limitados a menudo es conveniente utilizar tratamientos que son un poco menos eficaces cuando son mucho más baratos que el mejor del mercado.)

Pregunta cuatro: ¿Son las intervenciones razonables y viables en los contextos en los que es probable que se vayan a aplicar?

Un ensayo de investigación que compare una intervención poco clara e inasequible con otra tendrá poca influencia sobre la práctica médica. Se debe recordar que la práctica estándar actual (que puede consistir en «no hacer nada») debería ser casi invariablemente una de las alternativas comparadas. Demasiados ensayos de investigación analizan grupos de intervenciones que serían imposibles de implementar en un contexto ajeno a la investigación (p. ej., asumen que los médicos generales tendrán un sistema informático actualizado y se comprometerán a seguir un protocolo, que se dispone de infinito tiempo de personal de enfermería para realizar análisis de sangre o que los pacientes tomarán sus decisiones personales sobre el tratamiento basándose únicamente en el criterio de valoración primario del ensayo).

Pregunta cinco: ¿Qué método de análisis se utilizó, y era apropiado?

Esta decisión se puede resumir del siguiente modo (v. sección «Medición de los costes y beneficios de las intervenciones sanitarias»):

(a) Si las intervenciones produjeron resultados idénticos ⇒ análisis de minimización de costes.
(b) Si el resultado importante es unidimensional ⇒ análisis de coste-efectividad.
(c) Si el resultado importante es multidimensional ⇒ análisis de coste-utilidad.
(d) Si los resultados se pueden expresar de manera significativa en términos monetarios (es decir, si es posible sopesar la ecuación coste-beneficio para esta enfermedad frente a la ecuación coste-beneficio para otra enfermedad) ⇒ análisis de coste-beneficio.
(e) Si un análisis de coste-beneficio fuese apropiado por lo demás, pero los valores de preferencia otorgados a diferentes estados de salud están cuestionados o es probable que cambien ⇒ análisis de coste-consecuencias.

Pregunta seis: ¿Cómo se miden los costes y beneficios?

Volvamos por un momento a la sección «Medición de los costes y beneficios de las intervenciones sanitarias», donde esbocé algunos de los costes asociados con mi operación de apéndice. Ahora imaginemos un ejemplo más complejo: la rehabilitación de pacientes con ictus en sus propios domicilios con asistencia a un centro de día en comparación con una intervención alternativa estándar (rehabilitación en un hospital de larga estancia). El análisis económico debe tener en cuenta no sólo el tiempo de los distintos profesionales implicados, el tiempo de las secretarias y administrativos que ayudan a que el servicio funcione, así como el coste de la comida y los fármacos consumidos por los pacientes con ictus, sino también una parte del coste económico de construir el centro de día y el mantenimiento de un servicio de transporte hacia y desde él.

No hay reglas fijas para decidir qué costes deben incluirse. Si se va a calcular el «coste por caso» desde el principio, debe recordarse que hay que pagar por la calefacción, la iluminación, el personal de apoyo e incluso las facturas de los contables de la institución. En términos generales, estos «costes ocultos» se denominan gastos generales y suelen sumar un 30-60% adicional al coste de un proyecto. La tarea de costear operaciones y consultas externas en Reino Unido es más fácil de lo que solía ser porque actualmente estas actuaciones se venden y compran a un precio que refleja (o debería reflejar) todos los gastos generales implicados. Sin embargo, hay que tener en cuenta que los costes unitarios de las intervenciones sanitarias calculados en un país a menudo no guardan relación con los de la misma intervención en otros lugares, incluso cuando estos costes se expresan como porcentaje del PIB.

Los beneficios como la reincorporación más rápida al trabajo de un individuo en particular pueden medirse en términos del coste de emplear a esa persona con su salario habitual. Este enfoque tiene la consecuencia desafortunada y políticamente inaceptable de valorar más la salud de los profesionales que la de los trabajadores manuales, amas de casa o desempleados, y la de la mayoría de

raza blanca más que la de los grupos étnicos minoritarios (en general) peor pagados. Por lo tanto, podría ser preferible calcular el coste de los días de baja a partir del salario medio nacional.

En un análisis de coste-efectividad, los cambios del estado de salud se expresarán en unidades naturales (v. sección «Medición de los costes y beneficios de las intervenciones sanitarias»). Sin embargo, sólo porque las unidades sean naturales no significa automáticamente que sean adecuadas. Por ejemplo, el análisis económico del tratamiento de la úlcera péptica con dos fármacos diferentes podría medir el resultado como «proporción de úlceras cicatrizadas después de un tratamiento de 6 semanas». Los tratamientos podrían compararse según el coste por úlcera cicatrizada. Sin embargo, si las tasas de recidiva con los dos fármacos fuesen muy diferentes, el fármaco A podría considerarse falsamente más coste-eficaz que el fármaco B. Una mejor medida de resultado en este caso podría ser «el número de úlceras que se mantenían curadas al cabo de un año». En el análisis de coste-beneficio, donde el estado de salud se expresa en unidades de utilidad, como AVAC, si se fuese muy riguroso en la evaluación del artículo, habría que recordar cómo se obtuvieron las medidas de utilidad particulares usadas en el análisis (v. sección «Medición de los costes y beneficios de las intervenciones sanitarias»). En especial, habría que saber qué valores de preferencia de salud se utilizaron: los de los pacientes, los médicos, los economistas sanitarios o los del gobierno.

Pregunta siete: ¿Se tuvieron en cuenta los beneficios incrementales en lugar de los absolutos?

Esta pregunta se ilustra mejor con un ejemplo sencillo. Supongamos que el fármaco X, que cuesta 100 £ por tratamiento, cura a 10 de cada 20 pacientes. Su nuevo competidor, el fármaco Y, cuesta 120 £ por tratamiento y cura a 11 de cada 20 pacientes. El coste por caso curado con el fármaco X es de 200 £ (porque se han gastado 2.000 £ para curar a 10 personas) y el coste por caso curado con el medicamento Y es de 218 £ (porque se han gastado 2.400 £ para curar a 11 personas).

El coste *incremental* del fármaco Y, es decir, el coste adicional de curar al paciente adicional no es 18 £ sino 400 £ ya que ésta es la cantidad total adicional que se ha tenido que pagar para lograr un resultado por encima de lo que se habría logrado administrando a todos los pacientes el fármaco más barato. Este sorprendente ejemplo debe tenerse en cuenta la próxima vez que un representante farmacéutico intente convencernos de que su producto es «más eficaz y sólo ligeramente más caro».

Pregunta ocho: ¿Se ha dado prioridad al «aquí y ahora» respecto al futuro lejano?

Más vale pájaro en mano que ciento volando. En materia de salud, al igual que en términos económicos, se valora mucho más un beneficio en la actualidad que una promesa del mismo beneficio dentro de 5 años. Cuando los costes o beneficios de una intervención (o de la ausencia de la intervención) se producirán en algún momento del futuro, se debe realizar un descuento de su valor para reflejarlo. La cantidad real de descuento que se debe permitir para un

beneficio de salud futuro, frente a uno inmediato, es bastante arbitraria, pero en la mayoría de los análisis se utiliza una cifra de alrededor del 5% anual.

Pregunta nueve: ¿Se realizó un análisis de sensibilidad?

Supongamos que un análisis de coste-beneficio indica que la reparación de la hernia mediante cirugía ambulatoria cuesta 1.500 £ por AVAC, mientras que la reparación abierta tradicional, con su estancia hospitalaria asociada, cuesta 2.100 £ por AVAC. Con todo, si nos fijamos en cómo se realizaron los cálculos, nos sorprenderemos de lo barato que se ha tasado el equipo laparoscópico. Si se aumentase el precio de este equipo un 25%, ¿la cirugía ambulatoria seguiría teniendo este coste sorprendentemente más barato? Puede que sí, o puede que no.

El análisis de sensibilidad o la evaluación de los planteamientos hipotéticos: «¿qué pasaría sí…?» se describe en la sección «Validación de las pruebas diagnósticas frente a un patrón oro» del capítulo 9 en relación con los metaanálisis. En este caso se aplican exactamente los mismos principios: si el ajuste de las cifras para tener en cuenta toda la gama de posibles influencias proporciona una respuesta totalmente diferente, no se debe confiar demasiado en el análisis. Se puede consultar un buen ejemplo de un análisis de sensibilidad sobre un tema de interés tanto científico como político en un artículo sobre el coste-efectividad del tratamiento con estatinas en personas con diferentes niveles de riesgo basal de enfermedades cardiovasculares[11].

Pregunta diez: ¿Se han usado en exceso las escalas agregadas «resumidas»?

En la sección «Medición de los costes y beneficios de las intervenciones sanitarias» introduje el concepto del análisis de coste-consecuencias, en el que el lector del artículo puede añadir sus propios valores a las diferentes utilidades. En la práctica, ésta es una forma inusual de presentar un análisis económico y lo más habitual es que el lector se enfrente a un análisis de coste-utilidad o de coste-beneficio, que proporciona una puntuación compuesta expresada en unidades poco familiares que no se traducen fácilmente en los beneficios y perjuicios que el paciente puede esperar exactamente. La situación es análoga a la del padre al que se le dice: «el cociente intelectual de su hijo es de 115», cuando estaría mucho mejor informado si se le ofreciesen los datos desglosados: «su hijo puede leer, escribir, contar y dibujar bastante bien para su edad».

Conclusión

Espero que este capítulo haya demostrado que para realizar la evaluación crítica de un análisis económico es tan crucial responder a preguntas del tipo: «¿de dónde proceden estos números?» y «se han olvidado algunas cifras?» como comprobar que las propias sumas sean correctas. Aunque pocos artículos cumplen todos los criterios enumerados en la sección «Diez preguntas que se deben plantear acerca de un análisis económico» y que se resumen en el apéndice 1, después de leer el capítulo, el lector debería ser capaz de distinguir un análisis económico de calidad metodológica moderada o buena de otro que incluya cálculos banales sobre los

costes («el fármaco X cuesta menos que el fármaco Y; por lo tanto, es más coste-eficaz») en su sección de resultados o de discusión.

Bibliografía

1 DrummondM, Jefferson T. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ: British Medical Journal 1996;**313**(7052):275.
2 Jefferson T, Demicheli V, Mugford M. *Elementary economic evaluation in health care*. London: BMJ Books; 2000.
3 Donaldson C, Mitton C. *Priority setting toolkit: guide to the use of economics in healthcare decision making*. Oxford: JohnWiley & Sons; 2009.
4 McDowell I, Newell C, McDowell I. *Measuring health: a guide to rating scales and questionnaires*. New York: Oxford University Press; 2006.
5 Bradley C, Speight J. Patient perceptions of diabetes and diabetes therapy: assessing quality of life. Diabetes/Metabolism Research and Reviews 2002;**18**(S3):S64-9.
6 Bache I. Measuring quality of life for public policy: an idea whose time has come? Agenda-setting dynamics in the European Union. Journal of European Public Policy 2013;**20**(1):21-38.
7 Fairclough DL. *Design and analysis of quality of life studies in clinical trials*. Boca Raton: CRC Press; 2010.
8 Phillips D. *Quality of life: concept policy and practice*. Oxon: Routledge; 2012.
9 Young T, Yang Y, Brazier JE, et al. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. Quality of Life Research 2009;**18**(2):253-65.
10 Henderson C, KnappM, Fernández J-L, et al. Cost effectiveness of telehealth for patients with long termconditions (Whole Systems Demonstrator telehealth questionnaire study): nested economic evaluation in a pragmatic, cluster randomised controlled trial. BMJ: British Medical Journal 2013;346:f1035.
11 Jha V, Modi GK. Cardiovascular disease: the price of a QALY – cost-effectiveness of statins in CKD. Nature Reviews Nephrology 2013;**9**:377-9.
12 Herman WH, Edelstein SL, Ratner RE, et al. The 10-year cost-effectiveness of lifestyle intervention ormetformin for diabetes prevention: an intent-to-treat analysis of the DPP/DPPOS. Diabetes Care 2012;**35**(4):723-30.
13. Martin NK, Vickerman P, Miners A, et al. Cost-effectiveness of hepatitis C virus antiviral treatment for injection drug user populations. Hepatology 2012;**55**(1):49-57.
14 Saariniemi KM, Kuokkanen HO, Räsänen P, et al. The cost utility of reduction mammaplasty at medium-term follow-up: a prospective study, Journal of Plastic. Reconstructive & Aesthetic Surgery 2012;**65**(1):17-21.
15 Shahab L: *Cost-effectiveness of pharmacotherapy for smoking cessation*. London: National Centre for Smoking Cessation and Training (NCSCT), 2012 Available online http://www.ncsct.co.uk/usr/pub/B7_Cost-effectiveness _pharmacotherapy.pdf; accessed 5.11.13.
16 Coyle D, Coyle K, Cameron C, et al. Cost-effectiveness of new oral anticoagulants compared with warfarin in preventing stroke and other cardiovascular events in patients with atrial fibrillation. Value in Health 2013;**16**:498-506.
17 Frederix GW, Severens JL, Hövels AM, et al. Reviewing the cost-effectiveness of endocrine early breast cancer therapies: influence of differences in modeling methods on outcomes. Value in Health 2012;**15**(1):94-105.

18 Harris J. QALYfying the value of life. Journal ofMedical Ethics 1987;**13**(3):117-23.
19 Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. British Medical Bulletin 2010;**96**(1):5-21.
20 Gold MR, Stevenson D, Fryback DG. HALYS and QALYS and DALYS, Oh My: similarities and differences in summary measures of population health. Annual Review of Public Health 2002;**23**(1):115-34.

Capítulo 12 Artículos que van más allá de los números (investigación cualitativa)

¿Qué es la investigación cualitativa?

Hace veinticinco años, cuando asumí mi primer cargo de investigación, un colega que estaba cansado de su trabajo me aconsejó: «encuentra algo para medir y sigue midiendo hasta que tengas una caja llena de datos. Entonces deja de medir y empieza a escribir».

Yo le pregunté: «pero ¿qué debo medir?».

Su respuesta, cargada de cinismo fue: «eso da igual».

Este ejemplo verídico ilustra las limitaciones de una perspectiva exclusivamente cuantitativa (recuento y medición) en la investigación. El epidemiólogo Nick Black ha argumentado que un hallazgo o un resultado tiene más probabilidades de ser aceptado como un hecho si se cuantifica (se expresa en números) que en caso contrario[1]. La evidencia científica que respalda, por ejemplo, los «hechos» bien conocidos de que una pareja de cada 10 es infértil o que una persona de cada 10 es homosexual es escasa o nula. Sin embargo, según señala Black, la mayoría de nosotros aceptamos tan contentos y sin cuestionarlas estas afirmaciones simplistas, reduccionistas y descaradamente incorrectas siempre que contengan al menos un número.

Los investigadores cualitativos buscan una verdad más profunda. Su objetivo es «estudiar las cosas en su entorno natural», en un intento de dar sentido o interpretar fenómenos en términos de los significados que las personas les otorgan[2], y utilizan una «perspectiva holística que conserva las complejidades de la conducta humana»[2].

La investigación interpretativa o cualitativa ha sido durante años el territorio de los científicos sociales. En la actualidad, se reconoce cada vez más no sólo como un elemento complementario sino, en muchos casos, como un prerrequisito para la investigación cuantitativa con la que la mayoría de quienes nos hemos formado en las ciencias biomédicas está más familiarizada. Sin duda, la opinión de que los dos enfoques son mutuamente excluyentes se ha vuelto «poco científica» a su vez y en la actualidad está bastante de moda (sobre todo en los campos de la investigación sobre atención primaria y servicios de salud) afirmar que se está llevando

a cabo alguna investigación cualitativa y, desde que se publicó la primera edición de este libro, la investigación cualitativa se ha convertido incluso en la corriente principal del movimiento de la medicina basada en la evidencia[3,4] y, como se describe en el capítulo 7, se han producido importantes avances en la ciencia de la integración de la evidencia cualitativa y cuantitativa en el desarrollo y la evaluación de intervenciones complejas.

El difunto Dr. Cecil Helman, antropólogo y médico, me contó la siguiente historia para ilustrar la dicotomía cualitativo-cuantitativo. Un niño pequeño entra corriendo del jardín y dice emocionado: «Mami, las hojas se están cayendo de los árboles».

«Dime más», dice su madre.

«Bueno, se han caído cinco hojas en la primera hora y luego diez hojas cayeron en la segunda hora…»

Ese niño se convertirá en un investigador cuantitativo.

Un segundo hijo, cuando se le pidió que dijese más, podría haber respondido: «bueno, las hojas son grandes y planas, y en su mayoría de color amarillo o rojo, y parece que están cayendo de algunos árboles, pero no de otros. Y mamá, ¿por qué no cayeron hojas el mes pasado?»

Ese niño se convertirá en un investigador cualitativo.

Las preguntas como: «¿cuántos padres consultarían con su médico de familia cuando su hijo tenga febrícula?» o «¿qué proporción de los fumadores han intentado dejar de fumar?» necesitan claramente ser respondidas mediante métodos cuantitativos. Sin embargo, las preguntas como: «¿por qué los padres se preocupan tanto por la temperatura de sus hijos?» y «¿qué impide que las personas dejen de fumar?» no pueden y no deberían responderse precipitándose y midiendo el primer aspecto del problema que nosotros (los profanos en la materia) creamos que podría ser importante. En su lugar, tenemos que dedicar tiempo, escuchar lo que la gente tiene que decir y analizar las ideas y preocupaciones que tienen las propias personas. Después de un tiempo, será posible observar cómo aparece un patrón, que puede llevarnos a establecer nuestras observaciones de manera diferente. Podemos empezar con alguno de los métodos que se muestran en la tabla 12.1 y luego pasar a utilizar una selección de otros.

En el cuadro 12.1, que se reproduce con autorización del artículo introductorio de Nick Mays y Catherine Pope «Qualitative Research in Health Care»[5] se resumen (de un modo exagerado) las diferencias entre los enfoques cualitativo y cuantitativo de la investigación. En realidad, hay una gran cantidad de solapamiento entre ellos, cuya importancia cada vez se reconoce más[6].

Como se explica en la sección «Tres preguntas preliminares para orientarse» del capítulo 3, la investigación cuantitativa debe comenzar con una idea (por lo general, articulada en forma de hipótesis), que después, a través de la medición, genera datos y, por *deducción,* permite extraer una conclusión. La investigación cualitativa es diferente. Comienza con la intención de explorar un área en particular, recoge «datos» (p. ej., observaciones, entrevistas y documentos; incluso los correos electrónicos pueden considerarse datos cualitativos) y genera ideas e

Tabla 12.1 Ejemplos de métodos de investigación cualitativa

Etnografía (observación pasiva)	Observación sistemática de la conducta y conversación en entornos naturales
Etnografía (observación participante)	Observación en la que el investigador también ocupa un papel o parte en el entorno, además de observar
Entrevista semiestructurada	Conversación cara a cara (o telefónica) con el propósito de analizar asuntos o temas en detalle. Utiliza una amplia lista de preguntas o temas (denominada *guía temática*)
Entrevista narrativa	Entrevista realizada de una manera menos estructurada con el fin de conseguir una amplia historia del entrevistado (por lo general, una historia sobre la vida o la historia de cómo una enfermedad se ha desarrollado a lo largo del tiempo). El entrevistador se abstiene de estimular al entrevistado excepto para decir: «cuénteme más»
Grupos focales	Método de entrevista de grupo que incluye y utiliza explícitamente la interacción de grupo para generar datos
Análisis del discurso	Estudio detallado de las palabras, frases y formatos utilizados en contextos sociales particulares (incluye el estudio de conversaciones naturales, así como de materiales escritos, como documentos de políticas o actas de las reuniones)

Cuadro 12.1 Investigación cualitativa frente a cuantitativa: dicotomía exagerada (v. referencia 7)

	Cualitativa	Cuantitativa
Teoría social	Acción	Estructura
Métodos	Observación, entrevista	Experimento, estudio
Pregunta	¿Qué es X? (clasificación)	¿Cuántas X? (enumeración)
Razonamiento	Inductivo	Deductivo
Método de muestreo	Teórico	Estadístico
Fuerza	Validez	Fiabilidad

hipótesis a partir de estos datos en gran medida a través de lo que se denomina *razonamiento inductivo*[2]. La fuerza del enfoque cuantitativo reside en su *fiabilidad* (repetibilidad), es decir, las mismas mediciones deben obtener los mismos resultados una y otra vez. La fuerza de la investigación cualitativa radica en la *validez* (cercanía a la verdad), es decir, una buena investigación cualitativa, que utilice una selección de métodos de recopilación de datos, realmente debe llegar a la esencia de lo que está pasando en lugar de quedarse sólo en la superficie. Se ha afirmado

que la validez de los métodos cualitativos está mejorando en gran medida por el uso de la combinación de más de un método (v. tabla 12.1) (proceso denominado a veces *triangulación*), por parte del investigador que está pensando cuidadosamente sobre lo que está pasando y sobre el modo en el que su propia perspectiva podría influir en los datos (enfoque denominado *reflexividad*)[7], y, como aducirían algunas personas, por parte de más de un investigador que esté analizando los mismos datos de forma independiente (para demostrar la *fiabilidad interevaluadores*).

Desde que escribí la primera edición de este libro, la fiabilidad interevaluadores ha perdido credibilidad como medida de calidad en la investigación cualitativa. Los evaluadores de los artículos cualitativos tratan cada vez más de evaluar la competencia y la reflexividad de un investigador concreto en lugar de confirmar que los resultados fueron «verificados por otra persona». Este cambio es atribuible a dos ideas importantes. En primer lugar, en la mayor parte de la investigación cualitativa, una persona conoce los datos mucho mejor que cualquier otra, así que la idea de que dos cabezas piensan mejor que una simplemente no es cierta. Un investigador que se haya incorporado únicamente para verificar «temas» puede basarse mucho más en prejuicios y conjeturas personales que el trabajador de campo principal. Y en segundo lugar, con la tendencia a que más personas procedentes de entornos biomédicos realicen investigación cualitativa, no es infrecuente que dos investigadores no entrenados (o incluso un equipo completo de ellos) creen grupos focales o se lancen a evaluar las respuestas de texto libre de los cuestionarios. El «acuerdo» entre estos investigadores no sólo no corresponde a la calidad, sino que también es probable que los equipos procedentes de un entorno similar tengan sesgos parecidos, por lo que puntuaciones elevadas de fiabilidad interevaluadores pueden ser totalmente falsas.

Las personas sin conocimientos de investigación cualitativa suelen creer que equivale poco más o menos que a pasar el rato y ver las hojas caer. Detallar toda la literatura sobre cómo (y cómo no) actuar a la hora de observar, entrevistar y liderar un grupo focal, entre otras cosas, escapa al alcance de este libro. Sin embargo, existen métodos sofisticados para todas estas técnicas; quien esté interesado puede consultar la excelente serie del BMJ elaborada por Scott Reeves y cols. de Canadá[8-12].

Los métodos cualitativos realmente entran en acción a la hora de investigar un territorio desconocido, es decir, donde las variables de mayor interés no se conocen bien, están mal definidas y no se pueden controlar. En estas circunstancias, es posible que no se llegue a formular la hipótesis definitiva hasta que el estudio esté bastante avanzado. Sin embargo, es precisamente en estas circunstancias donde el investigador cualitativo debe asegurarse de que, desde el principio, ha delineado cuidadosamente un objetivo particular de la investigación y ha identificado algunas preguntas específicas para tratar de responder (v. Pregunta uno, sección «Evaluación de los artículos que describen la investigación cualitativa»). Los métodos de la investigación cualitativa permiten (y de hecho, requieren) la modificación de la pregunta de investigación a la luz de los resultados generados durante el proceso, técnica denominada *enfoque progresivo*[5]. (En cambio, como se

mostraba en la sección «¿Han planteado los autores correctamente el escenario?» del capítulo 5, curiosear los resultados provisionales de un estudio cuantitativo es estadísticamente inválido.)

El denominado enfoque *iterativo* (modificar los métodos de investigación y la hipótesis sobre la marcha) empleado por los investigadores cualitativos presenta una sensibilidad encomiable hacia la riqueza y la variabilidad del tema. En el pasado, dado que no se aceptaba la legitimidad de este enfoque, los críticos acusaban a los investigadores cualitativos de variar continuamente sus propias reglas de juego. Aunque estas críticas suelen ser erróneas, existe el peligro de que cuando unos investigadores ingenuos llevan a cabo una investigación cualitativa sin rigor, el enfoque «iterativo» será pasto de la confusión. Ésta es una de las razones por las que los investigadores cualitativos deben tomarse un tiempo de vez en cuando alejados de su trabajo de campo para la reflexión, planificación y consulta con colegas.

Evaluación de los artículos que describen la investigación cualitativa

La propia naturaleza de la investigación cualitativa hace que sea no estándar, no restringida y dependiente de la experiencia subjetiva tanto del investigador como del investigado. Analiza lo que se debe analizar y establece sus conclusiones en consonancia. Como se deduce de la sección anterior, la investigación cualitativa es una tarea interpretativa en profundidad, no un procedimiento técnico. Depende fundamentalmente de que un investigador competente y experimentado ejerza el tipo de habilidades y criterios que son difíciles, si no imposibles, de medir objetivamente. Por lo tanto, es discutible si se podría haber elaborado una lista de comprobación global para la evaluación crítica similar al *Users' Guides to the Medical Literature* para la investigación cuantitativa, aunque se han hecho intentos valerosos[3,4,10,13]. Algunas personas han argumentado que las listas de comprobación para la evaluación crítica pueden ir en detrimento de la calidad de la investigación en la investigación cualitativa porque fomentan un enfoque mecanicista y dirigido por protocolos[14].

Mi propio punto de vista y el de varias personas que han intentado o que están trabajando actualmente en esta precisa tarea es que tal vez esta lista de comprobación no sea tan exhaustiva o tan aplicable de forma universal como las diversas guías para evaluar la investigación cuantitativa, pero no cabe duda de que se pueden establecer algunas reglas básicas. Sin duda, Dixon-Woods y cols.[15] han llevado a cabo el mejor intento de ofrecer orientación (y también la mejor explicación de las incertidumbres e incógnitas). A continuación, se presenta una lista elaborada a partir de trabajos publicados que se citan en otros apartados del capítulo y también de las conversaciones que mantuve hace muchos años con el Dr. Rod Taylor, quien elaboró una de las primeras guías de evaluación crítica para los artículos cualitativos.

Pregunta uno: ¿El artículo describe un problema clínico importante abordado mediante una pregunta claramente formulada?

En la sección «Tres preguntas preliminares para orientarse» del capítulo 3 expliqué que una de las primeras cosas que se deben buscar en cualquier artículo de investigación es una explicación de por qué se llevó a cabo la investigación y qué pregunta específica abordaba. Los artículos cualitativos no son una excepción a esta regla: entrevistar u observar a la gente por amor al arte carece de cualquier utilidad científica. Los artículos que no pueden definir su tema de investigación con más precisión que «decidimos entrevistar a 20 pacientes con epilepsia» inspiran poca confianza en que los investigadores realmente sabían lo que estaban estudiando o por qué.

Es posible que se sienta mayor inclinación a leer el artículo si éste indica en su introducción algo así como: «la epilepsia es una enfermedad frecuente y potencialmente incapacitante, y una proporción significativa de pacientes sigue teniendo crisis pese a tomar la medicación. Se sabe que los fármacos anticomiciales tienen efectos secundarios desagradables y varios estudios han demostrado que una alta proporción de pacientes no toman sus pastillas con regularidad. Por lo tanto, decidimos analizar las creencias de los pacientes sobre la epilepsia y sus razones percibidas para no tomar su medicación».

Como expliqué en la sección «¿Qué es la investigación cualitativa?», la naturaleza iterativa de dicha investigación hace que la pregunta de investigación definitiva no se pueda plantear claramente al comienzo del estudio, pero sin duda debería haberse formulado en el momento de escribir el artículo.

Pregunta dos: ¿Fue apropiado usar un enfoque cualitativo?

Si el objetivo de la investigación era explorar, interpretar u obtener una comprensión más profunda de un problema clínico particular, es casi seguro que utilizar los métodos cualitativos fue lo más adecuado. Sin embargo, si la investigación buscaba alcanzar algún otro objetivo (como determinar la incidencia de una enfermedad o la frecuencia de una reacción adversa a un fármaco, poner a prueba una hipótesis de causa-efecto o demostrar que un fármaco tiene una mejor relación riesgo-beneficio que otro), los métodos cualitativos son claramente inapropiados. Si se cree que un estudio de casos y controles, un estudio de cohortes o un ensayo aleatorizado habría sido más adecuado para la pregunta de investigación planteada en el artículo que los métodos cualitativos que se utilizaron en realidad, se podría comparar esa pregunta con los ejemplos de la sección «Ensayos controlados aleatorizados» para confirmar la sospecha.

Pregunta tres: ¿Cómo se seleccionaron (a) el contexto y (b) los sujetos?

Volvamos a observar el cuadro 12.1, donde se comparan los métodos de muestreo *estadísticos* de la investigación cuantitativa con los métodos *teóricos* de la investigación cualitativa. A continuación se explicará lo que esto significa. En capítulos anteriores, en especial en la sección «¿Qué pacientes incluye el estudio?» (cap. 4), destaqué la importancia, en la investigación cuantitativa, de

garantizar que se recluta una muestra verdaderamente aleatoria de participantes. Una muestra aleatoria garantizará que los resultados reflejan, en promedio, el estado de la población de la que se extrajo la muestra.

Sin embargo, en la investigación cualitativa no nos interesa una visión «promedio» de una población de pacientes. El objetivo es comprender en profundidad la experiencia de individuos o grupos específicos, por lo que se deben buscar deliberadamente individuos o grupos que reúnan los requisitos. Si, por ejemplo, el objetivo fuese estudiar la experiencia de las mujeres cuando dieron a luz en el hospital, estaría perfectamente justificado desviarse un poco y buscar a mujeres que hubiesen tenido diversas experiencias de parto, como un parto inducido, una cesárea de urgencia, un parto atendido por un estudiante de medicina, un aborto involuntario tardío, etcétera.

Es posible que también quisiéramos seleccionar algunas mujeres que hubiesen recibido asistencia prenatal combinada por parte de un obstetra y de su médico de cabecera, y algunas mujeres que hubiesen sido atendidas por matronas de la comunidad durante todo el embarazo. En este ejemplo podría ser especialmente instructivo encontrar a mujeres que hubiesen recibido asistencia de médicos varones a pesar de que ésta sería una situación relativamente infrecuente. Por último, podría optarse por estudiar a las pacientes que dieron a luz en el contexto de una gran maternidad, moderna y de alta tecnología, así como a algunas que lo hicieron en un pequeño hospital comunitario. Por supuesto, todas estas especificaciones nos darán muestras «sesgadas», pero eso es exactamente lo que queremos.

Hay que tener cuidado con la investigación cualitativa donde la muestra se haya seleccionado (o parezca haberse seleccionado) únicamente en función de la comodidad. En el ejemplo mencionado previamente, aprovechar la primera docena de pacientes que pasasen por el paritorio más cercano sería la forma más fácil de conseguir las entrevistas, pero la información obtenida tal vez fuese mucho menos útil.

Pregunta cuatro: ¿Cuál fue la perspectiva del investigador, y se ha tenido en cuenta?

Dado que la investigación cualitativa se basa necesariamente en la experiencia de la vida real, un artículo que describa este tipo de investigación no debe ser «criticado» simplemente porque los investigadores hayan planteado una perspectiva cultural particular o una implicación personal con los participantes de la investigación. Más bien al contrario: deberían ser felicitados por hacer precisamente eso. Es importante reconocer que es imposible abolir o controlar completamente el sesgo del observador en la investigación cualitativa. Esto es más obvio cuando se utiliza la observación participante (v. tabla 12.1), pero también se cumple para otras formas de recogida y de análisis de datos.

Si, por ejemplo, la investigación se refiriese a la experiencia de adultos con asma que viven en una vivienda húmeda con mucha gente y al efecto percibido de este entorno en su salud, los datos generados por técnicas como grupos focales

o entrevistas semiestructuradas probablemente estarían muy influenciados por lo que el *entrevistador* cree acerca de este tema y por el hecho de que estuviese contratado por una clínica de neumología, por el departamento de trabajo social de la autoridad local o por un grupo de presión medioambiental. Sin embargo, dado que es inconcebible que las entrevistas pudiesen haber sido realizadas por alguien que no tuviese ningún punto de vista ni perspectiva ideológica o cultural, lo más que se puede exigir a los investigadores es que describan en detalle cuál es su punto de vista para que los resultados puedan interpretarse en consecuencia.

Por este motivo los investigadores cualitativos suelen preferir escribir sus trabajos en primera persona («entrevisté a los participantes» en lugar de «se entrevistó a los participantes») porque así se reflejan de forma explícita el papel y la influencia del investigador.

Pregunta cinco: ¿Qué métodos utilizó el investigador para la recogida de datos y se describen con suficiente detalle?

Una vez pasé 2 años llevando a cabo investigación experimental altamente cuantitativa basada en el laboratorio en la que se dedicaban alrededor de 15 horas todas las semanas al llenado o vaciado de tubos de ensayo. Había una manera estándar de llenar los tubos, una manera estándar de centrifugarlos e incluso una manera estándar de lavarlos. Cuando finalmente publiqué mi investigación, unas 900 horas de aquella tarea ingrata quedaron resumidas en una sola frase: se midieron los niveles séricos de ruibarbo de los «pacientes» según el método descrito por Bloggs y Bloggs (indicando la referencia del artículo de Bloggs y Bloggs sobre el modo de medir el ruibarbo sérico).

Ahora dedico gran parte de mi tiempo a la investigación cualitativa y puedo confirmar que es infinitamente más divertida. Mis colegas de investigación y yo hemos dedicado unos 15 años al análisis de las creencias, esperanzas, temores y actitudes de los pacientes diabéticos de los grupos étnicos minoritarios en el East End de Londres (comenzamos con los bangladesíes británicos y ampliamos el trabajo a otros grupos étnicos del sur de Asia, y luego a otros). Tuvimos que desarrollar, por ejemplo, una forma válida de traducir simultáneamente y transcribir las entrevistas que se realizaron en Sylheti, un complejo dialecto bengalí que no tiene forma escrita. Observamos que las actitudes de los participantes parecen estar fuertemente influenciadas por la presencia en la consulta de algunos de sus familiares, por lo que nos las ingeniamos para entrevistar a algunos pacientes tanto en presencia como en ausencia de esos familiares clave.

Podría seguir describiendo los métodos que ideamos para abordar este tema de investigación en particular, pero probablemente ya he dejado clara mi posición: la sección de métodos de un artículo cualitativo a menudo no se puede escribir de forma abreviada ni pasarse por alto haciendo referencia a las técnicas de investigación de otra persona. Es posible que deba ser larga y detallada, ya que está contando una historia única sin la cual los resultados no se pueden interpretar. Al igual que sucede con la estrategia de muestreo, no hay reglas fijas

sobre qué detalles exactamente deberían incluirse en esta sección del artículo. Simplemente habría que preguntar: «¿se ofrece suficiente información sobre los métodos utilizados?». En caso afirmativo, hay que utilizar el sentido común para valorar si «¿son estos métodos una manera sensata y adecuada de abordar la pregunta de investigación?».

Pregunta seis: ¿Qué métodos utilizó el investigador para analizar los datos y qué medidas de control de calidad se aplicaron?

La sección de análisis de datos de un artículo de investigación cualitativa ofrece al investigador o investigadores la oportunidad de demostrar la diferencia entre lo que tiene sentido y lo que no. Después de haber recopilado una gran cantidad de transcripciones de entrevistas o notas de campo, el verdadero investigador cualitativo apenas ha comenzado su tarea. No basta con hojear los textos en busca de citas interesantes que apoyen una teoría en especial. El investigador debe encontrar una manera *sistemática* de analizar sus datos y, en particular, debe tratar de detectar e interpretar elementos de los datos que parezcan contradecir o cuestionar las teorías derivadas de la mayor parte de la información. Cathy Pope y Sue Ziebland publicaron uno de los mejores artículos cortos sobre el análisis de datos cualitativos en el *British Medical Journal* hace unos años y los lectores que sean nuevos en este campo y quieran saber por dónde empezar deberían consultarlo[16]. Quien quiera conocer el libro definitivo sobre investigación cualitativa, donde se describen varios enfoques diferentes para el análisis, debería leer la magnífica obra editada por Denzin y Lincoln[2].

La forma más frecuente con gran diferencia de analizar los datos cualitativos que suelen recopilarse en la investigación biomédica es el *análisis temático.* En él, los investigadores analizan escritos de texto libre, elaboran una lista de temas generales y asignan categorías de codificación a cada uno. Por ejemplo, un «tema» podría ser el conocimiento que tienen los pacientes sobre su enfermedad y dentro de este tema podría haber códigos, como incluir «causas transmisibles», «causas sobrenaturales», «causas debidas a la propia conducta», etcétera. Hay que tener en cuenta que estos códigos no se corresponden con una taxonomía biomédica convencional («genética», «infecciosa», «metabólica», etc.) ya que la finalidad de la investigación es analizar la taxonomía de los entrevistados, tanto si se está de acuerdo con ella como si no. El análisis temático suele abordarse mediante la elaboración de una matriz o tabla, con una nueva columna para cada tema y una nueva fila para cada «caso» (p. ej., una transcripción de la entrevista), y cortando y pegando segmentos relevantes del texto en cada celda[13]. Otro tipo de análisis temático es el método comparativo constante, en el cual cada nuevo dato se compara con el resumen emergente de todos los elementos anteriores, lo que permite el refinamiento secuencial de una teoría emergente[17].

En la actualidad, es muy frecuente que el análisis de datos cualitativos se realice con la ayuda de un programa informático como ATLAS-TI o NVIVO, lo que facilita en gran medida manejar grandes conjuntos de datos. Las

afirmaciones hechas por todos los entrevistados sobre un tema en particular pueden compararse entre sí y se pueden realizar comparaciones sofisticadas, como: «¿las personas que hicieron la afirmación A también tienden a hacer la afirmación B»? De todas formas, hay que recordar que un programa informático cualitativo no analiza los datos por sí sólo y ningún programa cuantitativo, como el SPSS, puede indicar al investigador qué prueba estadística debería aplicar en cada caso. Aunque una frase como «los datos se analizaron utilizando NVIVO» podría parecer impresionante, si los datos utilizados son de mala calidad, los resultados del análisis también lo serán. Se puede realizar un análisis excelente de datos cualitativos en el que los textos impresos de entrevistas (por ejemplo) se marcan con rotuladores y (por ejemplo) se utiliza el método comparativo constante de forma manual en vez de un método informático.

A la hora de escribir sobre investigación cualitativa, a menudo es difícil demostrar cómo se llevó a cabo el control de calidad. Como se ha indicado en la sección anterior, sólo porque más de un investigador haya analizado los datos no se garantiza *necesariamente* el rigor. De hecho, los investigadores que nunca están en desacuerdo en sus juicios subjetivos (¿un párrafo específico del relato de un paciente indica en realidad «ansiedad», «alienación» o «confianza»?) probablemente nunca analizan lo suficiente sus propias interpretaciones. La esencia de la calidad en tales circunstancias tiene más que ver con el nivel de diálogo crítico entre los investigadores y en *cómo* se expusieron y se resolvieron los desacuerdos. Al analizar los datos de mis primeras investigaciones sobre las creencias acerca de la salud de los bangladesíes británicos con diabetes, por ejemplo, tres investigadores de mi equipo analizamos una transcripción escrita de una entrevista y asignamos códigos a afirmaciones particulares[18]. A continuación, comparamos nuestras decisiones y argumentamos (a veces acaloradamente) nuestros desacuerdos. Nuestro análisis reveló diferencias en la interpretación de ciertas afirmaciones que no pudimos resolver por completo. Por ejemplo, nunca llegamos a un acuerdo sobre lo que significa el término *ejercicio* en este grupo étnico. Esto no quiere decir que alguno de nosotros estuviese «equivocado», sino que había *ambigüedades* inherentes en los datos. Tal vez, por ejemplo, los entrevistados de esta muestra estuviesen confundidos acerca de lo que significa el término *ejercicio* y los beneficios que proporciona a las personas con diabetes.

Pregunta siete: ¿Los resultados son creíbles y, si es así, son clínicamente importantes?

Es evidente que no se puede evaluar la credibilidad de los resultados cualitativos mediante la precisión y exactitud de los aparatos de medición, ni por su significación con intervalos de confianza y los números necesarios a tratar. La herramienta fundamental para determinar si los resultados son razonables y creíbles, y si son importantes en la práctica, es el simple sentido común.

Un aspecto importante que se debe comprobar en la sección de resultados es si los autores citan datos reales. Afirmaciones como: «los médicos generales no suelen reconocer la utilidad de la evaluación anual» serían más creíbles si se reprodujeran una o dos citas textuales de los entrevistados para ilustrarlas. Los resultados deben poderse verificar de forma independiente y objetiva (p. ej., incluyendo segmentos más largos de texto en un apéndice o un recurso en línea), y todas las citas y ejemplos deberían indexarse para que se pudiese determinar su correspondencia con un entrevistado y fuente de datos identificables.

Pregunta ocho: ¿Qué conclusiones se extrajeron y están justificadas por los resultados?

Un artículo de investigación cuantitativa, presentado según el formato estándar Introducción, Métodos, Resultados y Discusión (IMRYD) (v. sección «La ciencia de criticar los artículos» del capítulo 3), debería distinguir claramente los resultados del estudio (por lo general, un conjunto de números) de la interpretación de esos resultados. El lector no debería tener dificultades para distinguir lo que *encontraron* los investigadores de lo que ellos creen que *significa*. Sin embargo, en la investigación cualitativa esta distinción pocas veces es posible, ya que los resultados son, por definición, una interpretación de los datos.

Por lo tanto, al evaluar la validez de la investigación cualitativa, es necesario preguntar si la interpretación de los datos coincide con el sentido común y que la perspectiva personal, profesional y cultural del investigador se expresen de forma explícita para que el lector pueda evaluar la «lente» a través de la cual el investigador ha realizado el trabajo de campo, el análisis y la interpretación. Esto puede resultar difícil porque el lenguaje que usamos para describir las cosas tiende a otorgar significados y motivos que los propios participantes pueden no compartir. Compárense, por ejemplo, estas dos frases: «tres mujeres fueron al pozo a por agua» y «tres mujeres se encontraron en el pozo y cada una llevaba una jarra».

Se está convirtiendo en un cliché que las conclusiones de los estudios cualitativos, al igual que las de cualquier investigación, deberían estar «basadas en la evidencia», es decir, que deben extraerse a partir de lo que los investigadores encontraron en la literatura. Mays y Pope[5] proponen tres preguntas útiles para determinar si las conclusiones de un estudio cualitativo son válidas:

- ¿En qué medida este análisis explica por qué las personas se comportan de la manera en que lo hacen?
- ¿Qué grado de comprensibilidad tendría esta explicación para un participante reflexivo en este ámbito?
- ¿En qué medida la explicación coincide con lo que ya sabemos?

Pregunta nueve: ¿Los resultados del estudio son transferibles a otros contextos?

Una de las críticas más frecuentes que se hace a la investigación cualitativa es que los resultados de cualquier estudio cualitativo sólo son válidos para el

contexto limitado en el que se obtuvieron. En realidad, esto puede suceder tanto con la investigación cualitativa como con la cuantitativa. Volvamos por un momento al ejemplo de las experiencias de parto en las mujeres que he descrito en la pregunta tres. Una muestra de conveniencia de la primera docena de mujeres que diesen a luz proporcionaría poca más información aparte de las experiencias recogidas de estas 12 mujeres. Una muestra *intencional,* como la que se describe en la pregunta tres, ampliaría la transferibilidad de las conclusiones a las mujeres que tienen una amplia gama de experiencias de parto. Sin embargo, si se realizan ajustes reiterados en el marco de muestreo a medida que se desarrolla el estudio de investigación, los investigadores podrían elaborar una muestra teórica y evaluar las nuevas teorías a medida que vayan surgiendo. Por ejemplo (y se trata de un ejemplo inventado), los investigadores podrían descubrir que las mujeres con mayor nivel educativo parecen tener experiencias más traumáticas desde el punto de vista psicológico que las mujeres con menor nivel educativo. Esto podría dar lugar a una nueva hipótesis sobre las expectativas de las mujeres (cuanto mayor es el nivel educativo de la mujer, mayores son sus expectativas de una «experiencia de parto perfecta»), que a su vez motivaría un cambio en la estrategia de muestreo intencional (ahora se querría encontrar los extremos del nivel educativo materno), etcétera. Cuanto más se dirija la investigación por este tipo de análisis de datos progresivo de tipo centrado y reiterado, más probable es que sus resultados sean transferibles fuera de la propia muestra.

Conclusión

Los médicos han otorgado tradicionalmente un alto valor a los datos basados en números, que en realidad pueden ser engañosos, reduccionistas e irrelevantes para los problemas reales. La creciente popularidad de la investigación cualitativa en las ciencias biomédicas se debe en gran parte a que los métodos cuantitativos o bien no proporcionan respuestas, u ofrecen respuestas incorrectas a preguntas importantes, tanto en la asistencia clínica como en la prestación de servicios. Quien todavía piense que la investigación cualitativa es de segunda categoría por ser una ciencia subjetiva debería saber que se ha quedado desfasado respecto a la evidencia.

En 1993, Catherine Pope y Nicky Britten presentaron en una conferencia un artículo titulado «Barriers to qualitative methods in the medical mindset» (Barreras a los métodos cualitativos en la mentalidad médica), donde mostraron una colección de cartas de rechazo de revistas biomédicas[19]. Las cartas revelaban una ignorancia sorprendente de la metodología cualitativa por parte de los revisores. Dicho de otro modo, las personas que habían rechazado los artículos a menudo parecían ser incapaces de distinguir una buena investigación cualitativa de una de mala calidad.

Irónicamente, en la actualidad se publican trabajos cualitativos de mala calidad con bastante frecuencia en algunas revistas médicas, que parecen haber experimentado un cambio de actitud de su política editorial desde la exposición de Pope y Britten sobre la «mentalidad médica». Por lo tanto, espero que las preguntas

enumeradas anteriormente y las referencias posteriores ayuden a los revisores de ambos bandos: los que siguen rechazando trabajos cualitativos por las razones equivocadas y los que se han subido al carro cualitativo y están aceptando tales artículos por razones incorrectas. Sin embargo, se debe tener en cuenta que la valoración crítica de la investigación cualitativa es una ciencia relativamente poco desarrollada y las preguntas planteadas en este capítulo todavía se están perfeccionando.

Bibliografía

1 Black N. Whywe need qualitative research. Journal of Epidemiology and Community Health 1994;**48**(5):425-6.
2 Denzin NK, Lincoln YS. *The SAGE handbook of qualitative research*. London: Sage; 2011.
3 Giacomini MK, Cook DJ. Users' guides to themedical literature XXIII. Qualitative research in health care A. Are the results of the study valid? JAMA:The Journal of the American Medical Association 2000;**284**(3):357-62.
4 Giacomini MK, Cook D. Users' guides to the medical literature: XXIII. Qualitative research in health care B. What are the results and how do they help me care for my patients? JAMA: The Journal of the American Medical Association 2000;**284**:478-82.
5 Mays N, Pope C. Qualitative research in health care: assessing quality in qualitative research. BMJ: British Medical Journal 2000;**320**(7226):50.
6 Dixon-Woods M, Agarwal S, Young B, et al. *Integrative approaches to qualitative and quantitative evidence*. London: Health Development Agency; 2004.
7 Gilgun JF. Reflexivity and qualitative research. Current Issues in Qualitative Research 2010;**1**(2):1-8.
8 Reeves S, Albert M, Kuper A, et al. Qualitative research: why use theories in qualitative research? BMJ: British Medical Journal 2008;**337**(7670):631-4.
9 Lingard L, Albert M, Levinson W. Grounded theory, mixed methods, and action research. BMJ: British Medical Journal 2008;**337**(aug07_3):a567-667.
10 Kuper A, Lingard L, Levinson W. Critically appraising qualitative research. British Medical Journal 2008;**337**:a1035.
11 Kuper A, Reeves S, Levinson W. Qualitative research: an introduction to reading and appraising qualitative research. BMJ: British Medical Journal 2008;**337**(7666):404-7.
12 Reeves S, Kuper A, Hodges BD. Qualitative researchmethodologies: ethnography. BMJ: British Medical Journal 2008;**337**:a1020.
13 Spencer L, Britain G. *Quality in qualitative evaluation: a framework for assessing research evidence*. Cabinet Office, London: Government Chief Social Researcher's Office; 2003.
14 Barbour RS. Checklists for improving rigour in qualitative research a case of the tail wagging the dog? BMJ: British Medical Journal 2001;**322**(7294):1115.
15 Dixon-Woods M, Shaw RL, Agarwal S, et al. The problemof appraising qualitative research. Quality and Safety in Health Care 2004;**13**(3):223-5.
16 Pope C, Ziebland S, Mays N. Qualitative research in health care: analysing qualitative data. BMJ: British Medical Journal 2000;**320**(7227):114.
17 Glaser BG. The constant comparative method of qualitative analysis. Social Problems 1965;**12**(4):436-45.

18 Greenhalgh T, Helman C, Chowdhury AM. Health beliefs and folk models of diabetes in British Bangladeshis: a qualitative study. BMJ: British Medical Journal 1998;**316**(7136):978-83.
19 Pope C, Britten N. The quality of rejection: barriers to qualitative methods in the medical mindset. Paper presented at BSA Medical Sociology Group annual conference 1993.

Capítulo 13 Artículos que describen investigaciones basadas en cuestionarios

El auge imparable de las investigaciones basadas en cuestionarios

¿Cuándo y dónde fue la última vez que tuvo que rellenar un cuestionario? Nos los entregan en casa o aparecen en nuestros casilleros en el trabajo, vienen como adjuntos del correo electrónico y los encontramos en la sala de espera del dentista. Los niños los traen a casa de la escuela y no es infrecuente que nos entreguen uno acompañando la cuenta del restaurante. Recientemente conocí a alguien en una fiesta que se describió a sí mismo como un «asaltante del cuestionario», pues su trabajo consistía en parar a la gente por la calle y anotar sus respuestas a una serie de preguntas acerca de sus ingresos, gustos, preferencias comerciales y mil cosas más.

Este capítulo se basa en una serie de artículos que edité para el *British Medical Journal,* escritos por un equipo dirigido por mi colega Petra Boynton[1-3]. Petra me ha enseñado mucho acerca de esta técnica de investigación ampliamente utilizada, incluido el hecho de que es probable que haya más investigaciones basadas en cuestionarios de mala calidad en la literatura que casi cualquier otro diseño de estudio. Mientras que se necesita un laboratorio para hacer un mal trabajo de laboratorio y un suministro de fármacos para hacer una mala investigación farmacéutica, todo lo que se requiere para realizar una mala investigación basada en cuestionarios es escribir una lista de preguntas, fotocopiarla y pedir a algunas personas que las contesten. Por lo tanto, es un poco extraño que las *Users' Guides to the Medical Literature,* que son muy exhaustivas en otros aspectos, publicadas en el *Journal of the American Medical Association,* no incluyan (hasta donde yo sé) un artículo sobre estudios basados en cuestionarios.

Los cuestionarios suelen considerase un medio «objetivo» de recogida de información sobre los conocimientos, las creencias, actitudes y conductas de las personas[4,5] ¿Les gusta a nuestros pacientes nuestro horario de apertura? ¿Qué piensan los adolescentes de una campaña local de lucha contra las drogas (y ha cambiado sus actitudes)? ¿Cuánto saben las enfermeras sobre el tratamiento del asma? ¿Qué proporción de la población se considera a sí misma como homosexual

o bisexual? ¿Por qué los médicos no aprovechan el máximo potencial de los ordenadores? Es probable que a partir de estos ejemplos pueda apreciarse que los cuestionarios permiten buscar datos tanto cuantitativos (al *x* por ciento de las personas les gustan nuestros servicios) como cualitativos (las personas que usan nuestros servicios tienen experiencias *xyz*). Dicho de otro modo, los cuestionarios no son un «método cuantitativo» o un «método cualitativo», sino una herramienta para recoger varios tipos distintos de datos, dependiendo de la pregunta planteada en cada ítem y el formato en el que se espera que los encuestados las respondan.

En el capítulo anterior ya indiqué que si los datos utilizados son de mala calidad, los resultados del estudio también lo serán, para puntualizar que los instrumentos mal estructurados proporcionan datos de mala calidad, conclusiones engañosas y recomendaciones confusas. Todo esto alcanza su máximo grado en la investigación basada en cuestionarios. Aunque ahora se dispone de forma generalizada de una orientación clara sobre el diseño y la forma de presentar los ensayos controlados aleatorizados (ECA) y las revisiones sistemáticas (v. el apartado acerca de la lista de comprobación CONSORT en el cap. 6 y las listas de comprobación QUORUM y PRISMA en el cap. 9), no existe un marco comparable para la investigación basada en cuestionarios, aunque parece que hay uno en fase de desarrollo. Tal vez por esta razón, a pesar de la gran cantidad de orientación detallada que existe en la literatura especializada[4,5], los errores metodológicos elementales son frecuentes en la investigación basada en cuestionarios realizada por profesionales sanitarios[1-3].

Antes de pasar a la valoración crítica, dedicaré unas líneas a aclarar la terminología. Un cuestionario es una forma de instrumento psicométrico, es decir, está diseñado para medir formalmente un aspecto de la psicología humana. A veces nos referimos a los cuestionarios como «instrumentos». Las preguntas de un cuestionario se denominan en ocasiones *ítems*. Un ítem es la unidad más pequeña del cuestionario que se puntúa de forma individual. Puede constar de una propuesta o indicación («señale cuál de las siguientes respuestas corresponde a su propia opinión») y luego cinco opciones posibles. También puede ser una simple respuesta de sí/no o verdadero/falso.

Diez preguntas que deben plantearse sobre un artículo que describa un estudio basado en cuestionarios

Pregunta uno: ¿Cuál era la pregunta de investigación, y era el cuestionario adecuado para contestarla?

En la sección «La ciencia de criticar los artículos» del capítulo 3 describo tres preguntas preliminares para empezar a evaluar cualquier artículo. La primera de ellas era: «¿cuál fue la pregunta de investigación y por qué era necesario el estudio?». Ésta es una pregunta inicial idónea para los estudios basados en cuestionarios porque (como se explica en la sección previa) los investigadores inexpertos a menudo se embarcan en la investigación basada en cuestionarios sin aclarar por qué lo están haciendo o lo que quieren averiguar. Además, la

gente a menudo decide utilizar un cuestionario para estudios que necesitan un método totalmente diferente. En ocasiones, un cuestionario puede ser apropiado, pero sólo si se utiliza en un estudio de metodología mixta (p. ej., para ampliar y cuantificar los resultados de una fase exploratoria inicial). En la tabla 13.1 se presentan algunos ejemplos reales basados en los artículos que Petra Boynton y yo recopilamos a partir de la literatura publicada y que ofrecemos a los participantes en los cursos que hemos impartido.

El uso de un cuestionario previamente validado y publicado tiene muchas ventajas para los investigadores. El equipo de investigación ahorrará tiempo y recursos, podrá comparar sus propios resultados con los de otros estudios, sólo necesitará ofrecer unos detalles resumidos del instrumento al escribir su trabajo y no tendrá que haber pasado por un proceso exhaustivo de validación del instrumento. Por desgracia, los investigadores inexpertos (más frecuentemente, los estudiantes que realizan una tesis) tienden a olvidar buscar a fondo en la literatura para encontrar un instrumento «prefabricado» adecuado, y a menudo no conocen las técnicas de validación formales (v. el texto posterior). A pesar de que la mayoría de estos estudios serán rechazados por los editores de revistas, una preocupante proporción acaba publicada en la literatura.

En la investigación sobre servicios sanitarios se utilizan cada vez más los cuestionarios «prefabricados» estándar diseñados expresamente para obtener datos que se puedan comparar entre los estudios. Por ejemplo, los ensayos clínicos suelen incluir instrumentos estándar para medir lo que saben los pacientes sobre una enfermedad[6], la satisfacción con los servicios[7] o la calidad de vida (CDV) relacionada con la salud[8,9]. La validez (v. más adelante) de este enfoque depende esencialmente de si el tipo y el rango de respuestas cerradas (es decir, la lista de posibles respuestas entre las que se pide a las personas que elijan) refleja toda la gama de percepciones y sentimientos que las personas de los diferentes marcos de muestreo potenciales podrían tener en realidad.

Pregunta dos: ¿El cuestionario utilizado en el estudio era válido y fiable?

Un cuestionario válido mide lo que dice medir. En realidad, muchos no logran hacerlo. Por ejemplo, un cuestionario autocumplimentado cuya función sea medir la ingesta de alimentos de las personas quizá no sea válido porque en realidad mide lo que *dicen* que han comido, pero no lo que *realmente* han comido[10]. Del mismo modo, se ha demostrado que los cuestionarios que preguntan a los médicos generales cómo manejan determinados cuadros clínicos difieren significativamente de la práctica clínica real[11]. Se debe tener en cuenta que un instrumento desarrollado en una época, país o contexto cultural diferente puede no ser una medida válida en el grupo que se está estudiando. A continuación se presenta un ejemplo peculiar. El ítem «A menudo voy a fiestas homosexuales» era una medida válida del nivel de sociabilidad de la persona en Reino Unido en la década de 1950, pero la redacción tiene una connotación muy diferente en la actualidad[1]. Los lectores que estén interesados en la medición de la CDV mediante cuestionarios pueden consultar un artículo sobre la controversia acerca de la validez de dichos instrumentos cuando se utilizan fuera del contexto en el que se desarrollaron[12].

Tabla 13.1 Ejemplos de preguntas de investigación para las cuales es posible que un cuestionario *no* sea el diseño más apropiado

Área de investigación amplia	Ejemplo de preguntas de investigación	¿Por qué un cuestionario NO es el método más apropiado?	¿Qué método(s) debería(n) utilizarse en su lugar?
Carga de enfermedad	¿Cuál es la prevalencia del asma en escolares?	Un niño puede tener asma, pero el progenitor no saberlo; los progenitores pueden creer incorrectamente que su hijo tiene asma, o pueden no ofrecer información que se considere estigmatizante	Estudio transversal que utiliza criterios diagnósticos estandarizados y/o un análisis sistemático de las historias clínicas
Conducta profesional	¿Cómo tratan el lumbago los médicos generales?	Lo que los médicos dicen no es lo mismo que lo que hacen en realidad, sobre todo cuando creen que su práctica está siendo evaluada por otros	Observación directa o grabación en vídeo de las consultas; uso de pacientes simulados, o análisis sistemático de las historias clínicas
Estilo de vida relacionado con la salud	¿Qué proporción de personas en los estudios sobre dejar de fumar consigue dejarlo?	La proporción de personas que de verdad lo dejan es menor que la de quienes dicen haberlo dejado. Se observa un patrón similar en estudios sobre opciones dietéticas, ejercicio y otros factores del estilo de vida	Prueba diagnóstica considerada el patrón oro (en este ejemplo, cotinina urinaria o salival)
Evaluación de necesidades en grupos con «necesidades especiales»	¿Cuáles son las necesidades no cubiertas de los refugiados y de los solicitantes de asilo respecto a la asistencia sanitaria y servicios sociales?	Es probable que un cuestionario refleje las ideas preconcebidas de los investigadores (p. ej., puede tomar como su punto de partida los servicios existentes y/o las necesidades de grupos más «visibles») y no centrarse en áreas importantes de necesidades	Rango de métodos cualitativos de análisis diseñados para elaborar una imagen detallada del problema. Por ejemplo, entrevistas semiestructuradas a usuarios, profesionales sanitarios y voluntarios; grupos focales, así como estudios en profundidad de eventos críticos

Los cuestionarios fiables ofrecen resultados constantes a partir de muestras repetidas y de diferentes investigadores a lo largo del tiempo[4,5]. Las disparidades en los resultados obtenidos a partir de un cuestionario fiable se deben a las diferencias entre los participantes y no a las contradicciones en cómo se entienden los ítems o cómo los diferentes observadores interpretan las respuestas. Un cuestionario estandarizado es aquel que se escribe y se utiliza de un modo estrictamente establecido, por lo que a todos los participantes se les hacen exactamente las mismas preguntas en un formato idéntico y las respuestas se registran de manera uniforme. La estandarización de una medida aumenta su fiabilidad. Por ejemplo, en el censo de Reino Unido *(General Household Survey)* de 2011, se realizó una serie de preguntas de forma bastante mecánica porque el entrevistador fue entrenado para utilizar el instrumento de una manera altamente estandarizada con el fin de aumentar la fiabilidad. A menudo es difícil determinar a partir de un artículo publicado la vehemencia con la que los investigadores trataron de lograr la estandarización, pero es posible que indiquen las cifras de fiabilidad interevaluadores.

Pregunta tres: ¿Qué aspecto tenía el cuestionario, y era adecuado para la población diana?

Con el término «aspecto» me refiero a dos cosas: forma y contenido. La forma tiene que ver con cuestiones como el número de páginas que tenía, si era visualmente atractivo (o desagradable), cuánto tiempo se tardaba en completar, la terminología utilizada, etcétera. Éstas no son cuestiones menores. Un cuestionario que abarque 30 páginas, incluya interminables párrafos de jerga científica y contenga preguntas que un encuestado puede encontrar ofensivas no se rellenará adecuadamente, por lo que los resultados de un estudio carecerán de sentido[2].

El contenido tiene que ver con los ítems reales. ¿Las preguntas tienen sentido y podrían entenderlas los participantes de la muestra? ¿Había preguntas ambiguas o demasiado complicadas? ¿Se evitaban las palabras equívocas ambiguas, como «frecuentemente», «con regularidad», «habitualmente», «por lo general», «muchos», «algunos» y «casi nunca»? ¿Los ítems eran «abiertos» (los encuestados pueden escribir lo que quieran) o «cerrados» (los encuestados deben escoger de una lista de opciones) y, en este último caso, estaban representadas todas las respuestas posibles? Los diseños cerrados permiten a los investigadores obtener datos agregados rápidamente, pero el rango de posibles respuestas es fijado por los investigadores, no por los encuestados, por lo que la diversidad de las respuestas es mucho más baja[13]. Algunos de los encuestados (los propensos a contestar afirmativamente) tienden a estar de acuerdo con las afirmaciones en lugar de estar en desacuerdo. Por esta razón, los investigadores no deben presentar sus ítems de manera que «muy de acuerdo» siempre se relacione con la misma actitud amplia. Por ejemplo, en una escala de satisfacción del paciente, si una pregunta es «mi médico de cabecera en general trata de ayudarme», otra pregunta debería formularse de forma negativa, por ejemplo, «los recepcionistas *no suelen ser corteses*».

Pregunta cuatro: ¿Eran claras las instrucciones?

Quien alguna vez haya tenido que rellenar un cuestionario y se haya perdido a la mitad (o haya descubierto que no se sabe dónde hay que enviarlo una vez cumplimentado) sabrá que las instrucciones contribuyen decisivamente a la validez del instrumento. Entre ellas se incluyen:

- Una explicación de cuál es el tema del estudio y cuál es el propósito general de la investigación.
- La garantía de anonimato y confidencialidad, así como la confirmación de que la persona puede dejar de rellenar el cuestionario en cualquier momento sin tener que dar una razón.
- Datos de contacto claros y precisos de a quién dirigirse para obtener más información.
- Instrucciones sobre lo que se debe enviar y un sello de correos si se trata de un cuestionario postal.
- Instrucciones adecuadas sobre cómo completar cada ítem, con ejemplos cuando sea necesario.
- Cualquier encarte (p. ej., folleto), regalo (p. ej., bono descuento para un libro) u honorarios si son parte del protocolo.

Es improbable que estos aspectos del estudio se enumeren en el artículo publicado, pero pueden incluirse en un apéndice y, si no, se debería poder obtener la información de los autores.

Pregunta cinco: ¿Se realizó una prueba piloto adecuada del cuestionario?

Los cuestionarios a menudo fracasan porque los participantes no los entienden, no pueden completarlos, se aburren o se sienten ofendidos por ellos, o no les gusta su presentación. Aunque los amigos y compañeros de trabajo pueden ayudar a comprobar la ortografía, la gramática y el diseño, no pueden predecir de forma fiable las reacciones emocionales o las dificultades de comprensión de otros grupos. Por este motivo, se debe realizar una prueba piloto de todos los cuestionarios (tanto de nuevo desarrollo como «prefabricados») con participantes que sean representativos de la muestra de estudio definitiva que se va a evaluar, por ejemplo, cuánto tiempo se tarda en completar el instrumento, si algún ítem no se comprende o si la gente se aburre o se confunde a la mitad. Hay tres preguntas específicas que se deben hacer: i) ¿cuáles fueron las características de los participantes en quienes se realizó la prueba piloto del instrumento, ii) ¿*cómo* se realizó la prueba piloto; qué detalles se ofrecen? y iii) ¿*de qué forma* se modificó el instrumento definitivo como resultado de la prueba piloto?

Pregunta seis: ¿Cómo era la muestra?

Quien haya leído los capítulos anteriores sabrá que una muestra sesgada o no representativa dará lugar a resultados engañosos y a conclusiones inseguras. Al evaluar un estudio basado en cuestionarios, es importante preguntar cuál fue el marco de muestreo para el estudio definitivo (intencional, aleatorio y bola de nieve) y también si era lo suficientemente amplio y representativo.

A continuación, se indican los principales tipos de muestras para un estudio basado en cuestionarios (tabla 13.2):

- *Muestra aleatoria:* se identifica un grupo diana y se invita a participar a una selección aleatoria de las personas de ese grupo. Por ejemplo, se puede utilizar un ordenador para seleccionar al azar una muestra del 25% de pacientes de un registro de diabetes.
- *Muestra aleatoria estratificada:* es igual que la muestra aleatoria, pero el grupo diana se estratifica primero de acuerdo con una(s) característica(s) particular(es); por ejemplo, las personas con diabetes que reciben insulina, pastillas y dieta. El muestreo aleatorio se lleva a cabo de forma separada para estos distintos subgrupos.
- *Muestra de bola de nieve:* se identifica un pequeño grupo de participantes y luego se pide que «inviten a un amigo» a completar el cuestionario. A este grupo se le pide a su vez que designen a otra persona y así sucesivamente.
- *Muestra de oportunidad:* por lo general, por razones pragmáticas, a las primeras personas que parecen reunir los criterios se les pide que cumplimenten el cuestionario. Esto puede ocurrir, por ejemplo, en una concurrida consulta de medicina general cuando se pide a todos los pacientes que acuden un día en particular que rellenen una encuesta acerca de la comodidad del horario de apertura. Sin embargo, esta muestra presenta un sesgo claro porque quienes crean que dicho horario es incómodo no estarán allí para empezar. Este ejemplo debería recordarnos que estas muestras de oportunidad (a veces denominadas de conveniencia) pocas veces o nunca tienen una justificación científica.
- *Muestra sistemáticamente sesgada:* supongamos que queremos, por ejemplo, evaluar el grado de satisfacción de los pacientes con su médico de cabecera, y ya sabemos gracias al estudio piloto que el 80% de las personas con mayor nivel socioeconómico completará el cuestionario, pero sólo el 60% de los de menor nivel socioeconómico lo harán. Se podría realizar un sobremuestreo de este último grupo para garantizar que el conjunto de datos refleja la composición socioeconómica de la población práctica. (Idealmente, si se hace esto, también habrá que demostrar que las personas que se negaron a rellenar el cuestionario no diferían en las características clave respecto a quienes sí lo completaron.)

También es importante tener en cuenta si el instrumento era adecuado para todos los participantes y los participantes potenciales. En particular, debe determinarse si se tuvo en cuenta el rango probable que tenía la muestra en cuanto a capacidades físicas e intelectuales, lenguaje y alfabetización, comprensión de números o de escalas y de amenaza percibida de las preguntas o del encuestador.

Pregunta siete: ¿Cómo se aplicó el cuestionario, y fue la tasa de respuesta adecuada?

La sección de métodos de un artículo que describe un estudio basado en cuestionarios debe incluir detalles de tres aspectos de su aplicación: i) ¿cómo se distribuyó el cuestionario (p. ej., por correo postal, cara a cara o electrónicamente)?, ii) ¿cómo se cumplimentó el cuestionario (p. ej., autocumplimentación o con ayuda de un investigador)? y iii) ¿se describieron de forma completa las tasas

Tabla 13.2 Tipos de marco de muestreo para la investigación basada en cuestionarios

Tipo de muestra	Cómo funciona	Cuándo utilizarla
De oportunidad/al azar	Los participantes se seleccionan de un grupo que está disponible en el momento del estudio (p. ej., pacientes que acuden a una consulta una mañana particular)	Debería evitarse en la medida de lo posible
Aleatoria	Se identifica un grupo diana y se invita a participar a una selección aleatoria de ese grupo. Por ejemplo, se podría usar un ordenador para seleccionar de forma aleatoria una muestra del 25% de pacientes de un registro de diabéticos	Debe usarse en estudios donde se quiera reflejar la opinión promedio de una población
Aleatoria estratificada	Es igual que la muestra aleatoria, pero el grupo diana primero se estratifica según una(s) característica(s) particular(es). Por ejemplo, personas con diabetes que toman insulina, pastillas y siguen una dieta. El muestreo aleatorio se realiza por separado para estos subgrupos diferentes	Debe usarse cuando es probable que el grupo diana tenga diferencias sistemáticas por subgrupos
Cuota	Se identifican los participantes que encajan con la población más amplia (p. ej., en grupos, como clase social, sexo, edad, etc.). Los investigadores deben entrevistar a un número preestablecido de cada grupo (p. ej., x mujeres jóvenes de clase media)	Debe usarse en estudios donde se quiera reflejar unos resultados lo más representativos de la población más amplia como sea posible. Se usa con frecuencia en encuestas de opiniones políticas, entre otras
Bola de nieve	Se recluta a los pacientes y luego se les pide que identifiquen a otras personas similares para participar en la investigación	Es útil cuando se trabaje con grupos a los cuales es difícil llegar (p. ej., madres lesbianas)

de respuesta, incluidos los detalles de los participantes que eran inadecuados para la investigación o que se negaron a participar? y ¿se han comentado todos los posibles sesgos de respuesta?

El *British Medical Journal* no suele publicar un artículo que describa un estudio basado en cuestionarios si menos del 70% de la gente encuestada completó el cuestionario correctamente. Se han realizado varios estudios de investigación sobre el modo de aumentar la tasa de respuesta a un estudio basado en cuestionarios. En resumen, se ha demostrado que los siguientes factores aumentan las tasas de respuesta[3]:

- El cuestionario está diseñado con claridad y tiene un formato simple.
- Ofrece incentivos o premios a los participantes a cambio de su cumplimentación.
- Se ha realizado una prueba piloto exhaustiva y se ha evaluado.
- Los participantes reciben una notificación previa sobre el estudio, con una invitación personalizada.
- El objetivo del estudio y los medios de cumplimentar el cuestionario se explican claramente.
- Un investigador está disponible para responder preguntas y recoger el cuestionario cumplimentado.
- Si se utiliza un cuestionario postal, se incluye un sello de correos.
- Los participantes se sienten una parte implicada en el estudio.
- Las preguntas están redactadas de una manera que mantiene la atención de los participantes.
- El cuestionario tiene una orientación y propósito claros, y es conciso.
- El cuestionario es visualmente atractivo.

Otra cosa que debe buscarse respecto a las tasas de respuesta es una tabla en el artículo que compare las características de las personas que respondieron con las de las personas a quienes se entregó el cuestionario, pero que se negaron a cumplimentarlo. Si hubiera diferencias sistemáticas (en vez de aleatorias) entre estos grupos, los resultados de la encuesta no serían generalizables a la población de la cual se extrajeron los respondedores. Los respondedores a las encuestas realizadas en la calle, por ejemplo, suelen ser mayores que la media (tal vez porque tienen menos prisa) y es menos probable que sean de una minoría étnica (quizá porque algunos de estos últimos son incapaces de hablar el lenguaje del investigador con fluidez). Además, si los autores del estudio han demostrado que los no respondedores son bastante similares a los respondedores, habría que preocuparse menos por la generalizabilidad, incluso si las tasas de respuesta fuesen menores de lo que sería deseable.

Pregunta ocho: ¿Cómo se analizaron los datos?

El análisis de los datos de un cuestionario es una ciencia sofisticada. Quien esté interesado en aprender las técnicas formales puede consultar varios libros de texto excelentes sobre métodos de investigación social[4,5]. Quien sólo esté interesado en completar una lista de comprobación sobre un estudio basado en cuestionarios publicado, debe contemplar los siguientes aspectos del estudio. En primer lugar, ¿qué tipo de análisis a grandes rasgos se llevó

a cabo, y fue apropiado? En especial, ¿se utilizaron las pruebas estadísticas correctas para las respuestas cuantitativas, y/o se usó un método reconocible de análisis cualitativo (v. sección «Medición de los costes y beneficios de las intervenciones sanitarias» del capítulo 11) para preguntas abiertas? Aporta tranquilidad (aunque no es en absoluto una prueba perfecta) saber que uno de los autores del artículo es un estadístico. Además, como se afirmó en el capítulo 5, si se han usado pruebas estadísticas de las que nunca se ha oído hablar, probablemente debería sospecharse que hay gato encerrado. La gran mayoría de los datos del cuestionario se puede analizar mediante pruebas estadísticas de uso común, como chi cuadrado, Spearman, correlación de Pearson, etcétera. El error más frecuente de todos en la investigación basada en cuestionarios es no utilizar ninguna prueba estadística en absoluto y no se necesita un doctorado en estadística para detectar esta trampa.

También hay que comprobar que no hay evidencia de «dragado de datos», es decir, ¿los autores simplemente han introducido sus datos en un ordenador, han realizado cientos de pruebas y luego han ideado una hipótesis plausible que concuerde con algo que resulta ser «significativo»? Dicho de otro modo, todos los análisis deberían estar motivados por una hipótesis, es decir, la hipótesis se debe pensar primero y luego se debe realizar el análisis, y no viceversa.

Pregunta nueve: ¿Cuáles fueron los principales resultados?

Hay que tener en cuenta primero cuáles fueron los hallazgos globales y si se describieron todos los datos relevantes. ¿Son los resultados cuantitativos definitivos (estadísticamente significativos), y también se han descrito los resultados *no significativos* relevantes? Es posible que sea igual de importante descubrir, por ejemplo, que la confianza autonotificada de los médicos generales a la hora de tratar la diabetes *no* se correlaciona con sus conocimientos acerca de la enfermedad que descubrir que sí había una correlación. Por esta razón, un estudio basado en cuestionarios que sólo comente las asociaciones estadísticas positivas presenta sesgos internos.

Otra cuestión importante es si los resultados cualitativos se han interpretado de forma adecuada (p. ej., utilizando un marco teórico explícito) y si se han justificado y contextualizado debidamente todas las citas (en lugar de haberse escogido selectivamente para adornar el artículo). Aconsejo al lector que vuelva al capítulo 6 («Artículos que describen ensayos de tratamientos farmacológicos y otras intervenciones sencillas») para recordar los trucos utilizados por los responsables de marketing sin escrúpulos para exagerar los hallazgos. Se deben revisar cuidadosamente los gráficos (sobre todo los puntos de corte de los ejes) y las tablas de datos.

Pregunta diez: ¿Cuáles son las principales conclusiones?

Ésta es una pregunta de sentido común. ¿Qué significan realmente los resultados, y han establecido los investigadores una relación adecuada entre los datos y sus conclusiones? ¿Los resultados se han ubicado dentro del conjunto más amplio de conocimientos sobre el tema (especialmente cualquier estudio similar u opuesto que utilice el mismo instrumento)? ¿Han reconocido los autores

las limitaciones de su estudio y han redactado sus discusión a la luz de ellas? (p. ej., si la muestra era pequeña o la tasa de respuesta baja, ¿recomendaron más estudios para confirmar los resultados preliminares?). Por último, ¿están las recomendaciones plenamente justificadas por los resultados? Por ejemplo, si se ha realizado un estudio pequeño, a nivel muy local, no se deberían sugerir cambios en la política nacional como resultado. Quien sea nuevo en la valoración crítica puede encontrar dificultades a la hora de realizar estos juicios y la mejor manera de aprender es participar en los debates de las revistas (ya sea cara a cara o en línea), donde un grupo de personas comparten sus reacciones respecto a un artículo escogido.

En conclusión, cualquiera puede escribir una lista de preguntas y fotocopiarla, pero esto no significa que un conjunto de respuestas a estas preguntas constituya una investigación. El desarrollo, aplicación, análisis y presentación de estudios basados en cuestionarios es al menos tan difícil como los otros enfoques de investigación descritos en otros capítulos de este libro. Los investigadores que utilizan cuestionarios son un grupo heterogéneo y todavía no han consensuado un modelo de publicación estructurada comparable a CONSORT (ensayos controlados aleatorizados), QUORUM o PRISMA (revisiones sistemáticas) y AGREE (guías). Aunque ahora se dispone de varias herramientas estructuradas sugeridas, cada una diseñada para fines ligeramente distintos[14-16], una revisión de este tipo de herramientas encontró poco consenso y muchas preguntas sin respuesta[17]. Sospecho que a medida que estas guías se estandaricen y se usen más ampliamente, los artículos que describen la investigación basada en cuestionarios serán más coherentes y más fáciles de evaluar.

Bibliografía

1 Boynton PM, Wood GW, Greenhalgh T. A hands on guide to questionnaire research part three: reaching beyond the white middle classes. BMJ: British Medical Journal 2004;**328**(7453):1433-6.

2 Boynton PM, Greenhalgh T. A hands on guide to questionnaire research part one: selecting, designing, and developing your questionnaire. BMJ: British Medical Journal 2004;**328**(7451):1312-5.

3 Boynton PM. A hands on guide to questionnaire research part two: administering, analysing, and reporting your questionnaire. British Medical Journal 2004;**328**:1372-5.

4 Robson C. *Real world research: a resource for users of social research methods in applied settings*. Wiley: Chichester; 2011.

5 Bryman A. *Social research methods*. Oxford: Oxford University Press; 2012.

6 Dunn SM, Bryson JM, Hoskins PL, et al. Development of the diabetes knowledge (DKN) scales: forms DKNA, DKNB, and DKNC. Diabetes Care 1984;**7**(1):36-41.

7 Rahmqvist M. Bara A-C. Patient characteristics and quality dimensions related to patient satisfaction. International Journal for Quality in Health Care 2010;**22**(2):86-92.

8 Phillips D. *Quality of life: concept, policy and practice*. London: Routledge, 2012.

9 Bradley C, Speight J. Patient perceptions of diabetes and diabetes therapy: assessing quality of life. Diabetes/Metabolism Research and Reviews 2002;**18**(S3):S64-9.

10 Drewnowski A. Diet image: a new perspective on the food-frequency questionnaire. Nutrition Reviews 2001;**59**(11):370-2.
11 Adams AS, Soumerai SB, Lomas J, et al. Evidence of self-report bias in assessing adherence to guidelines. International Journal for Quality in Health Care 1999;**11**(3):187-92.
12 Gilbody S, House A, Sheldon T. Routine administration of Health Related Quality of Life (HRQoL) and needs assessment instruments to improve psychological outcome-a systematic review. Psychological Medicine 2002;**32**(8):1345-56.
13 Houtkoop-Steenstra H. *Interaction and the standardized survey interview: the living questionnaire*. Cambridge: Cambridge University Press, 2000.
14 Eysenbach G. Improving the quality ofWeb surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). Journal of Medical Internet Research 2004;**6**(3):e34.
15 Draugalis JR, Coons SJ, Plaza CM. Best practices for survey research reports: a synopsis for authors and reviewers. American Journal of Pharmaceutical Education 2008;**72**(1):11.
16 Kelley K, Clark B, Brown V, et al. Good practice in the conduct and reporting of survey research. International Journal for Quality in Health Care 2003;**15**(3):261-6.
17 Bennett C, Khangura S, Brehaut JC, et al. Reporting guidelines for survey research: an analysis of published guidance and reporting practices. PLoSMedicine 2011;**8**(8):e1001069.

Capítulo 14 Artículos que describen estudios sobre mejora de la calidad

¿Qué son los estudios sobre mejora de la calidad y cómo deberían investigarse?

El *British Medical Journal* (www.bmj.com) publica sobre todo artículos de investigación. *BMJ Quality and Safety* (http://qualitysafety.bmj.com), por su parte, publica principalmente descripciones de los esfuerzos para mejorar la calidad y seguridad de la asistencia sanitaria, a menudo en contextos reales, como plantas de hospital o consultas de medicina general[1]. Los estudiantes de medicina cuyos planes de estudios no incluyan temas de mejora de calidad pueden obviar este capítulo, cuyo contenido suele tratarse más en los cursos de posgrado. En cualquier caso, será necesario una vez que se hayan graduado y empiecen a ejercer su profesión.

Una forma fundamental de mejorar la calidad es aplicar los resultados de la investigación y realizar una asistencia más basada en la evidencia. Esto se comentará en el capítulo siguiente. Sin embargo, lograr una asistencia sanitaria de alta calidad y segura requiere algo más que una práctica basada en la evidencia. Pensemos, por ejemplo, en la última vez que nosotros o alguno de nuestros familiares estuvimos hospitalizados. Seguramente querríamos que se emplearan las pruebas diagnósticas más precisas (cap. 8), los fármacos (cap. 6) o las intervenciones no farmacológicas (cap. 7) más eficaces, y también que los médicos siguieran los planes de asistencia basados en la evidencia y guías (cap. 10) basadas en revisiones sistemáticas (cap. 9). Asimismo, si el hospital nos pidió que ayudáramos a evaluar el servicio, habríamos querido que se utilizase un cuestionario válido y fiable (cap. 13).

Sin embargo, ¿nos preocupamos también por cosas como cuánto habría que esperar para una cita ambulatoria y/o una intervención quirúrgica, las actitudes del personal, la claridad y exhaustividad de la información que se ofreció, el riesgo de contraer una infección (p. ej., cuando el personal no se

lavaba las manos constantemente) y la eficiencia general del centro? Si un miembro del personal cometió un error, ¿se nos explicó abiertamente y se ofreció una disculpa sin reservas? Y, en tal caso, ¿la organización tenía sistemas instaurados para aprender de lo que salió mal y asegurarse de que no vuelva a sucederle a otra persona? Una experiencia de asistencia sanitaria «de calidad» incluye todos estos aspectos y más. La ciencia de la mejora de la calidad obtiene su evidencia de muchas disciplinas diferentes, incluidas la investigación sobre la fabricación y el control del tráfico aéreo, así como de la medicina basada en la evidencia[2-4].

Mejorar la calidad y la seguridad en un área particular de la asistencia sanitaria suele implicar un proyecto complejo de al menos varios meses de duración, con aportaciones de diferentes miembros del personal (y cada vez más, también de los pacientes y sus representantes)[5]. Los directores del proyecto ayudan a todos los implicados a fijar una meta y a trabajar para conseguirla. El resultado del proyecto suele ser diverso: algunas cosas salen bien, otras cosas no tan bien y la iniciativa suele recogerse por escrito (en ocasiones) en forma de relato.

Desde hace varios años, el *BMJ* y el *BMJ Quality and Safety* distinguen entre los artículos de investigación (presentados según el formato de Introducción, Métodos, Resultados y Discusión, IMRD) y las publicaciones sobre mejora de calidad (presentados según el formato contexto, resumen del problema, medidas, proceso, análisis, estrategia para el cambio, efectos del cambio y pasos siguientes). Al hacer esta distinción, la investigación podría definirse como *una indagación sistemática y centrada en busca de verdades que sean transferibles más allá del contexto en el que se generaron*, mientras que la mejora de la calidad podría definirse como *un trabajo en tiempo real y en el mundo real llevado a cabo por los equipos que prestan servicios.*

Puede que el lector haya descubierto que existe una gran zona gris entre estas dos actividades. Parte de esta zona gris es la *investigación* sobre mejora de la calidad, es decir, la investigación aplicada orientada a la construcción de la base de evidencia sobre cómo se deberían realizar los estudios de mejora de la calidad. La investigación sobre la mejora de la calidad abarca una amplia gama de métodos, incluida la mayoría de los descritos en los demás capítulos. En particular, el *estudio con método mixto* incorpora tanto datos cuantitativos (p. ej., medidas de la prevalencia de una determinada enfermedad o problema) como datos cualitativos (p. ej., un análisis cuidadoso de los temas planteados en las cartas de queja o la observación participante del personal sobre sus propias obligaciones), escritos en forma de relato exhaustivo sobre lo que se hizo, por qué, cuándo, por quién y cuáles fueron las consecuencias. Si el artículo corresponde a una *investigación* verdadera sobre mejora de la calidad, debería incluir una conclusión que ofrezca lecciones transferibles para otros equipos en otros contextos[6,7].

Por cierto, mientras que un relato («anécdota») se considera de forma acertada como un diseño de estudio débil a la hora de, por ejemplo, evaluar la eficacia de un

fármaco, ese mismo formato (estudio de caso «organizacional») tiene ventajas específicas cuando se trata de reunir una gran cantidad de datos complejos y dar sentido a la misma, como sucede cuando una organización pretende mejorar su rendimiento[8].

Como es posible imaginar, la valoración crítica de la investigación sobre mejora de la calidad es un área particularmente difícil. A diferencia de los ensayos aleatorizados, no hay reglas fijas sobre cuál debe ser el «mejor» enfoque para una iniciativa de mejora de la calidad y es posible que se deba realizar una gran cantidad de juicios subjetivos sobre los métodos utilizados y la significación de los hallazgos. Aunque, como sucede con toda la evaluación crítica, cuantos más artículos se lean y evalúen, mejor se hará.

Al preparar la lista de preguntas de la sección siguiente, me he basado en gran medida en las guías SQUIRE (*Standards for QUality Improvement Reporting Excellence*, estándares para la excelencia en las publicaciones sobre mejora de la calidad), que son el equivalente de las guías CONSORT (*Consolidated Standards of Reporting Trials*, normas consolidadas para la publicación de ensayos clínicos), PRISMA (*Preferred Reporting Items for Systematic Reviews and Metaanalyses*, elementos de información preferidos para revisiones sistemáticas y metaanálisis), etcétera, para los estudios de mejora de la calidad[9]. Yo participé periféricamente en el desarrollo de estas guías y puedo confirmar que pasaron por múltiples revisiones antes de materializarse de forma impresa. Esto se debe a las dificultades *inherentes* de elaborar listas de comprobación estructuradas para evaluar estudios complejos y multifacéticos. Como se indica en el artículo del grupo de desarrollo de las guías SQUIRE (pág. 670):

> *A diferencia de las intervenciones conceptualmente nítidas y no ambiguas desde el punto de vista procedimental, como fármacos, pruebas y procedimientos, que afectan directamente a la biología de la enfermedad y son los objetos de estudio en la mayor parte de la investigación clínica, la mejora es esencialmente un proceso social. La mejora es una ciencia aplicada en lugar de una disciplina académica; su finalidad inmediata es cambiar el rendimiento humano en lugar de generar nuevos conocimientos generalizables y está impulsada principalmente por el aprendizaje experiencial. Al igual que otros procesos sociales, la mejora depende inherentemente del contexto. […] A pesar de que los métodos experimentales y cuasi-experimentales tradicionales son importantes para aprender si las intervenciones de mejora cambian el comportamiento, no proporcionan métodos apropiados y eficaces para abordar las preguntas pragmáticas cruciales, como por ejemplo: ¿cómo es el mecanismo de una intervención particular que funciona, para quién funciona y en qué circunstancias?*

Con estas advertencias en mente, a continuación veremos hasta dónde podemos llegar con una lista de comprobación para ayudar a dar sentido a los estudios de mejora de la calidad.

Diez preguntas que deben plantearse sobre un artículo que describa una iniciativa de mejora de la calidad

Después de haber desarrollado las siguientes preguntas, las apliqué a dos estudios sobre mejora de la calidad recientemente publicados. En mi opinión, me parecía que ambos tenían algunas características positivas, pero podrían haber sido mejores aún si se hubiesen publicado las guías SQUIRE cuando se escribieron los artículos. Es posible consultar estos artículos y seguir los ejemplos. Uno es un estudio de Verdú y cols.[10], de España, cuya finalidad era mejorar el tratamiento de la trombosis venosa profunda (TVP) en pacientes hospitalizados, y el otro es un estudio realizado por May y cols.[11], de EE.UU., cuyo objetivo era utilizar las visitas académicas (que Wikipedia define como la divulgación educativa sin una finalidad comercial, v. sección «Evidencia y marketing» del capítulo 6) para mejorar el tratamiento basado en la evidencia de la enfermedad crónica en un contexto de atención primaria.

Pregunta uno: ¿Cuál fue el contexto?

El «contexto» es el detalle local del entorno del mundo real en el que se realizó el trabajo. Más concretamente, uno de nuestros estudios de ejemplo tuvo lugar en España y el otro en Estados Unidos. Uno correspondía a atención secundaria y el otro a atención primaria. No es posible entender cómo se desarrollan estas diferentes iniciativas sin una cierta información sobre el país, el sistema de asistencia sanitaria y (a un nivel más local) los aspectos particulares históricos, culturales, económicos y micropolíticos de nuestro «caso».

Es útil, por ejemplo, no sólo saber que el estudio sobre visitas académicas de May y cols. se dirigió a médicos generales privados de EE.UU., sino también leer su breve descripción de la zona específica de Kentucky donde ejercían los médicos: «esta área tiene una demografía metropolitana regional que refleja una proporción considerable de la clase media de Estados Unidos (población, 260.512; renta media por hogar, 39.813 $; 19% de personas de raza no blanca; 13% por debajo de la línea de la pobreza; una ciudad, cinco comunidades rurales y cinco núcleos rurales históricamente con predominio de población de raza negra)»[11]. Por lo tanto, ésta era un área (clase media de EE.UU.) que, en general, no era ni especialmente rica ni especialmente pobre, que incluía zonas tanto urbanas como rurales, y con diversidad étnica, pero no excesiva.

Pregunta dos: ¿Cuál fue el objetivo del estudio?

No hace falta decir que el objetivo de un estudio de mejora de la calidad es mejorar la calidad. Quizás la mejor manera de plantear esta pregunta es: «¿cuál era el problema para el que una iniciativa de mejora de la calidad se consideraba una solución?».

En el ejemplo de Verdú y cols.[10] sobre TVP, los autores eran bastante sinceros al afirmar que el objetivo de su iniciativa de mejora de la calidad era ahorrar dinero. Más concretamente, trataban de reducir la duración de la hospitalización

de los pacientes (duración del «ingreso»). En el ejemplo sobre visita académica, un representante (visitador médico) se entrevistaba con los médicos para proporcionar una formación no sesgada y, en particular, para ofrecer directrices basadas en la evidencia para el tratamiento de la diabetes (primera visita) y del dolor crónico (segunda visita). El objetivo era ver si el modelo de visita académica, que ya se había mostrado en 1983 que mejoraba la práctica en ensayos de *investigación*[11], podría ser eficaz en el ambiente más confuso y menos predecible de la clase media de EE.UU.

Pregunta tres: ¿Cuál era el mecanismo por el cual los autores esperaban mejorar la calidad?

Esta pregunta acerca del CÓMO es fundamental. Volvamos por un momento a la sección «Diez preguntas que deben plantearse sobre un artículo que describa una intervención compleja» (v. cap. 7) sobre intervenciones complejas, donde la pregunta cuatro interrogaba por «*¿cuál era el mecanismo de acción teórico de la intervención?*». En realidad, se trata de la misma pregunta, aunque las iniciativas de mejora de calidad suelen tener unos límites difusos y no se debe esperar necesariamente que se identifique un «núcleo» claro para la intervención.

En el ejemplo de la vía clínica para la TVP, la iniciativa se basaba en el hecho de que si se desarrollaba una vía clínica integrada que incorporase todas las pruebas y tratamientos basados en la evidencia relevantes en el orden correcto, indicando quién era el responsable de cada paso y excluyendo cualquier actuación para la que hubiese evidencia de que no proporcionaba ningún beneficio, el personal la seguiría. Por consiguiente, el paciente podría pasar menos tiempo en el hospital y se emplearían menos procedimientos innecesarios. Además, los autores esperaban que, si se perfeccionaba la vía, también se reducirían los eventos adversos (como las hemorragias).

En el ejemplo sobre la visita académica, el «mecanismo» para modificar la conducta de prescripción de los médicos consistía en los principios de la influencia interpersonal y la persuasión a partir de los que la industria farmacéutica ha construido su estrategia de marketing (gran parte del cap. 6 se dedica a advertir de ellos). Los autores esperaban que si las guías se ofrecían y se explicaban en persona, aumentarían las probabilidades de seguirlas.

Pregunta cuatro: ¿La iniciativa de mejora de la calidad prevista estaba basada en la evidencia?

Algunas de las medidas destinadas a mejorar la calidad parecen una buena idea en teoría, pero en realidad no funcionan en la práctica. Tal vez el mejor ejemplo de esto sean las fusiones, es decir, la unión de dos organizaciones sanitarias pequeñas (p. ej., hospitales) con el fin de lograr ahorros de eficiencia, economías de escala, etcétera[12]. El equipo de Fulop demostró no sólo que estos ahorros pocas veces se materializan, sino que las organizaciones fusionadas a menudo se encuentran con problemas nuevos e imprevistos. En este ejemplo, además de no observar evidencia de que hubiese beneficios, se demostró que la iniciativa podría ser perjudicial.

En el ejemplo de la TVP, una revisión sistemática demuestra que, en general, en el contexto de la investigación, el desarrollo y la implementación de vías clínicas integrales (denominadas también *vías críticas*) *puede* reducir los costes y la duración de la estancia[13]. Del mismo modo, las revisiones sistemáticas han confirmado la eficacia de la visita académica en ensayos de investigación[14]. En nuestros dos ejemplos, después de contestar la pregunta ¿*puede* funcionar?, los autores plantearon una pregunta más específica y contextualizada: ¿funciona *aquí,* con *estos* pacientes y con *este* conjunto particular de *limitaciones* y *contingencias*?[15].

Pregunta cinco: ¿Cómo midieron el éxito los autores?, y ¿lo hicieron de forma razonable?

En un congreso reciente, visité una exposición de pósteres en los que los entusiastas de la medicina basada en la evidencia presentaban sus intentos de mejorar la calidad de un servicio. Algunos me impresionaron gratamente, pero me decepcionó bastante comprobar que, a menudo, los autores no habían medido formalmente el éxito de su iniciativa en absoluto o ni siquiera habían definido en qué habría consistido el éxito.

En nuestros dos ejemplos previos, los autores lo hicieron mejor. Verdú y cols. evaluaron su estudio sobre TVP en términos de seis resultados: duración de la hospitalización, coste de la asistencia hospitalaria y lo que ellos denominaron *indicadores de asistencia* (proporción de pacientes en cuya asistencia se siguió realmente la vía y proporción cuya duración de la hospitalización realmente se redujo de acuerdo con las recomendaciones de la vía, tasa de eventos adversos y nivel de satisfacción de los pacientes). Tomados en conjunto, estos resultados ofrecieron una buena indicación de si la iniciativa de mejora de la calidad había tenido éxito. Sin embargo, no era perfecta. Por ejemplo, el cuestionario de satisfacción no se había elaborado siguiendo adecuadamente los criterios de un buen estudio basado en cuestionarios descritos en el capítulo 13.

En el ejemplo de la visita académica, una buena medida del éxito de la iniciativa seguramente habría sido el grado en que los médicos siguieron las directrices o (mejor aún) el impacto sobre la salud y el bienestar de los pacientes, pero no se utilizaron estas medidas finales de resultados relevantes para los pacientes. En su lugar, la definición de «éxito» de los autores fue mucho más modesta: simplemente querían que sus representantes de guías basadas en la evidencia contactasen periódicamente con los médicos privados. Con ese fin, sus medidas de resultados incluían la proporción de médicos del área que accedieron a recibirles, la duración de la visita (despedirlos después de 45 segundos se consideraba una visita fallida), si el médico accedió a recibirlos en una segunda ocasión o subsiguiente y, en caso afirmativo, si podía encontrar fácilmente las guías proporcionadas en la primera visita.

Se podría alegar que estas medidas son el equivalente de los «criterios de valoración indirectos» que se describen en la sección «Criterios de valoración

indirectos» (cap. 6). Sin embargo, si se tiene en cuenta el contexto del mundo real (un grupo diana de médicos privados aislados desde los puntos de vista geográfico y profesional, y saturados de publicidad de la industria farmacéutica, para quienes la práctica basada en la evidencia no formaba parte de su forma de actuar), la posibilidad de ser recibidos es mucho mejor que nada. Sin embargo, al evaluar el artículo, debemos tener clara cuál era la modesta definición de éxito de los autores e interpretar las conclusiones en consecuencia.

Pregunta seis: ¿Cuánto detalle se dio sobre el proceso de cambio, y qué perspectivas se puede extraer de esto?

El aspecto más difícil de un intento de cambio suele estar en los detalles esenciales de la cuestión. En el ejemplo de la vía clínica de TVP, la sección de métodos era bastante corta y me dejó con ganas de más. Aunque me gustaron muchos aspectos del documento, me decepcionó esta brevísima descripción de lo que se hizo en realidad para *desarrollar* la vía: «después del diseño de la vía clínica, comenzamos el estudio». Sin embargo, ¿quién diseñó la vía y cómo? ¿Eran expertos en la práctica basada en la evidencia o personas que trabajaban en la primera línea asistencial? Lo ideal habría sido que hubiesen intervenido ambos, pero no lo sabemos. ¿Participaron sólo médicos en el proceso o también intervinieron enfermeras, farmacéuticos, pacientes y otros profesionales (como el gerente del hospital)? ¿Hubo desacuerdo sobre la evidencia o todos estaban de acuerdo en lo que se necesitaba? Cuanta más información sobre el *proceso* se pueda encontrar en el artículo, mejor se podrán interpretar los resultados tanto positivos como negativos.

En el ejemplo sobre visita académica, la sección de métodos es muy larga e incluye detalles sobre cómo se ha desarrollado el programa de la «visita», cómo se seleccionó y se formó a los representantes, cómo se eligió la muestra de médicos, cómo los representantes contactaron con los médicos, qué materiales de apoyo se utilizaron y cómo se estructuraron y se adaptaron las visitas a las necesidades y estilos de aprendizaje de los diferentes médicos. Tanto si estamos de acuerdo con sus medidas del éxito del proyecto como si no, es posible interpretar los resultados a la luz de esta información detallada sobre cómo lo llevaron a cabo.

La sección de métodos relativamente corta del ejemplo de la vía clínica de TVP puede haber sido víctima de los requisitos del número de palabras de la revista. Los autores resumen sus métodos con el fin de ser breves y, por lo tanto, dejan de lado todo el detalle cualitativo que habría permitido evaluar el *proceso* de mejora de la calidad, es decir, elaborar una panorámica detallada de lo que los autores hicieron realmente. Conocedores de este incentivo perverso, los autores de las guías SQUIRE remitieron una solicitud a los editores para que permitieran artículos «más largos»[9]. Un estudio sobre mejora de la calidad bien escrito podría abarcar una docena de páginas o más, y por lo general requeriría mucho más tiempo para leerse que, por ejemplo, una publicación bien escrita

sobre un ensayo aleatorizado. La tendencia creciente de las revistas a incluir material «extra» (precedido de «e-», indicando que está en línea) en un formato accesible en internet es muy alentadora y debería consultarse dicho material siempre que esté disponible.

Pregunta siete: ¿Cuáles fueron los principales hallazgos?

Para contestar a esta pregunta, hay que recordar lo que se ha contestado a la pregunta cinco, analizar los números (para los resultados cuantitativos) o los temas clave (para los datos cualitativos) y preguntar si fueron significativos y en qué modo. Al igual que en otros diseños de estudio, la «significación» de los estudios sobre mejora de la calidad es un concepto multifacético. Un cambio en un valor numérico puede ser clínicamente significativo sin ser estadísticamente significativo o viceversa (v. sección «Probabilidad y confianza»), y también puede ser susceptible de presentar diversos sesgos. Por ejemplo, en un estudio de tipo antes y después, el tiempo habrá variado entre las medidas «basales» y «postintervención», y diversos factores de confusión, como el clima económico, las actitudes de la opinión pública, la disponibilidad de fármacos o procedimientos particulares, la jurisprudencia pertinente, y la identidad del director ejecutivo, pueden haber cambiado. Los resultados cualitativos pueden ser especialmente vulnerables al efecto Hawthorne (el personal tiende a sentirse valorado y a trabajar más duro cuando se introduce cualquier cambio en las condiciones laborales destinadas a mejorar el rendimiento, tanto si tiene algún valor intrínseco como si no)[16].

En el ejemplo de la vía clínica sobre TVP, la estancia media se redujo en 2 días (una diferencia que fue estadísticamente significativa) y se consiguió un ahorro económico de varios cientos de euros por paciente. Además, se atendieron en realidad 40 de 42 pacientes elegibles mediante la nueva vía clínica (otros 18 pacientes con TVP no cumplían los criterios de inclusión) y el 62% de todos los pacientes alcanzaron el objetivo de reducción de la estancia hospitalaria. En general, 7 de 60 personas presentaron eventos adversos y en sólo uno de ellos se había seguido la vía clínica. Estas cifras, en su conjunto, no sólo reflejan que la iniciativa logró el objetivo de ahorrar dinero, sino que también ofrecen una clara indicación de la medida en la que se lograron los cambios previstos en el proceso de asistencia *y* nos recuerdan que muchos pacientes con TVP son lo que se denomina *excepciones,* es decir, el tratamiento basado en una vía estandarizada no se ajusta a sus necesidades.

En el ejemplo de la visita académica, los resultados muestran que de los 130 médicos del grupo diana, el 78% recibió, al menos, una visita y dicho grupo no presentaba diferencias en cuanto a las características demográficas (edad, sexo, formación en el extranjero o no) respecto a quienes rechazaron una visita. Sólo una persona se negó en redondo a recibir nuevas visitas, pero conseguir concertar otra visita fue muy complicado y las dificultades «se asociaron, sobre todo, con el hecho de convencer al personal administrativo de que el médico había aceptado nuevas visitas». Dicho de otro modo, a pesar de que el médico era (supuestamente) amable, los representantes tenían problemas

para superar la barrera de los recepcionistas, lo que seguramente sea un hallazgo cualitativo importante sobre el proceso de las visitas académicas, que no había sido observado en el diseño de los ensayos aleatorizados. La mitad de los médicos tenía a mano las directrices en la segunda visita (lo que implica que la otra mitad no las tenía). Sin embargo, el artículo también presentó algunos datos de resultados cuantitativos cuestionables, como que «alrededor del 90% de los médicos parecían interesados en los temas que se trataron», una observación que, además de ser totalmente subjetiva, es un efecto Hawthorne mientras no se demuestre lo contrario. En lugar de utilizar la técnica dudosa de tratar de cuantificar sus impresiones subjetivas, tal vez los autores deberían haberse ceñido a su medida de resultado principal (si los médicos les recibían o no) o haber llegado hasta el final y haber medido el cumplimiento de las guías.

Pregunta ocho: ¿Cuál fue la explicación para el éxito, el fracaso o la suerte dispar de la iniciativa? ¿Fue razonable?

Una vez más, las convenciones sobre la extensión de los artículos en revistas pueden hacer que esta sección sea frustrantemente corta. Lo ideal es que los autores hayan considerado sus resultados, repasado los factores contextuales que se identificaron en la pregunta uno y que hayan ofrecido una explicación plausible y razonada para sus resultados en términos de dichos factores, incluida una consideración de las explicaciones alternativas. Lo más frecuente es que las explicaciones sean breves y especulativas.

¿Por qué, por ejemplo, los representantes de la visita académica tenían dificultades para concertar segundas visitas con los médicos? Según los autores, la dificultad se debió a «la frecuente escasez de tiempo disponible para citas futuras y a factores operativos relacionados con la falta de financiación permanente para este servicio». Sin embargo, una explicación alternativa podría ser que el médico carecía de interés, pero no quería mostrarse intransigente, por lo que dijo a los recepcionistas que rechazasen a los representantes si volvían a acudir.

Al igual que en este ejemplo, la evaluación de las explicaciones que se ofrecen en un artículo sobre unos resultados decepcionantes de un proyecto de mejora de la calidad siempre es una cuestión de criterio. Nadie puede dar una lista de control que permita afirmar con el 100% de precisión: «*esta* explicación fue definitivamente creíble, mientras que *ese* aspecto definitivamente no lo era». En un estudio sobre mejora de calidad, los autores del artículo habrán contado una historia sobre lo que pasó y el lector tendrá que interpretar dicha historia usando sus conocimientos de medicina basada en la evidencia, sus conocimientos de las personas y las organizaciones, y su sentido común.

El artículo sobre la vía clínica para la TVP, aunque ofrece resultados muy positivos, también presenta una explicación realista de ellos: «el impacto real de las vías clínicas sobre la duración de la estancia es difícil de determinar debido a que estos estudios no aleatorizados y en parte retrospectivos podrían mostrar reducciones significativas de la estancia hospitalaria,

pero no se puede demostrar que la única causa de la reducción sea la vía clínica». ¡Faltaría más!

Pregunta nueve: A la luz de los resultados, ¿cuáles creen los autores que son los próximos pasos en el ciclo de mejora de la calidad a nivel local?

La calidad no es una estación a la que se llega, sino una manera de viajar. (Los lectores que quieran consultar una referencia sobre esta afirmación, pueden leer la obra de Pirsig[17] *Zen and the Art of Motorcycle Maintenance*.) Dicho de otro modo, la mejora de la calidad es un ciclo interminable: cuando se llega a una meta, se fija otra.

El equipo de la vía clínica para TVP se mostró satisfecho por haber reducido significativamente la duración de la estancia y pensó que la manera de mejorar más era garantizar que la vía clínica se modificase tan pronto como se dispusiera de nuevas pruebas y nuevas tecnologías. Otro planteamiento, que no mencionaron, pero que no tendría que esperar a una innovación, podría consistir en aplicar el enfoque de la vía clínica a una enfermedad médica o quirúrgica diferente.

El equipo de la visita académica decidió que su siguiente paso sería modificar el temario un poco para que, en lugar de abarcar dos temas no relacionados sobre áreas temáticas diferentes, se utilizase una «selección sensata de los temas secuenciales que permitiera reflejar de forma sutil los elementos clave del mensaje de las entrevistas anteriores (p. ej., tratamiento de la diabetes seguido de un programa de tratamiento de la hipertensión)». Es interesante observar que no tuvieron en cuenta abordar el problema del abandono (el 42% de los médicos no se prestaron a una segunda visita).

Pregunta diez: Según los autores, ¿cuáles eran las lecciones generalizables para otros equipos, y era esto razonable?

Al comienzo de este capítulo, argumenté que el rasgo distintivo de la investigación eran las lecciones generalizables para otros. No hay nada malo en mejorar de la calidad a nivel local sin tratar de obtener lecciones más amplias, pero si los autores han publicado su trabajo, a menudo afirman que otros deberían seguir su enfoque, o al menos determinados aspectos del mismo.

En el ejemplo de la vía clínica del tratamiento de la TVP, los autores no hacían afirmaciones sobre la transferibilidad de sus resultados. Su tamaño muestral era pequeño y ya se había demostrado que las vías clínicas acortan la estancia hospitalaria en otras afecciones comparables. Su motivo para la publicación parece que transmite el mensaje: «si nosotros pudimos hacerlo, los demás también pueden».

En el ejemplo de la visita académica se indicó que el hallazgo potencialmente transferible era que un enfoque centrado en toda la población para las visitas académicas (es decir, tratar de acceder a todos los médicos generales en una zona geográfica determinada) en comparación con dirigirse sólo a los voluntarios, puede «funcionar». Esta afirmación podría ser cierta, pero debido a que las medidas de resultado eran subjetivas y no tenían una relevancia directa para los pacientes, este estudio no logró demostrarla.

Conclusión

En este capítulo he intentado guiar al lector a través del modo de evaluar los artículos sobre estudios de mejora de la calidad. Como se ilustra por la cita que aparece al final de la sección «¿Qué son los estudios de mejora de la calidad y cómo deberían investigarse?», estos juicios son inherentemente difíciles de emitir y requieren integrar la evidencia y la información de múltiples fuentes. Por lo tanto, aunque los estudios de mejora de la calidad suelen ser pequeños, locales e incluso tienen cierta estrechez de miras, la evaluación crítica de este tipo de estudios a menudo es más complicada que la de un gran metaanálisis.

Bibliografía

1 Batalden PB, Davidoff F. What is "quality improvement" and how can it transform healthcare? Quality and Safety in Health Care 2007;**16**(1):2-3.
2 Marshall M. Applying quality improvement approaches to health care. BMJ: British Medical Journal 2009;339:b3411.
3 Miltner RS, Newsom JH, Mittman BS. The future of quality improvement research. Implementation Science 2013;**8**(Suppl 1):S9.
4 Vincent C, Batalden P, Davidoff F. Multidisciplinary centres for safety and quality improvement: learning from climate change science. BMJ Quality & Safety 2011;**20**(Suppl 1):i73-8.
5 Alexander JA, Hearld LR. The science of quality improvement implementation: developing capacity to make a difference. Medical Care 2011;**49**:S6-20.
6 Casarett D, Karlawish JH, Sugarman J. Determining when quality improvement initiatives should be considered research. JAMA: The Journal of the American Medical Association 2000;**283**(17):2275-80.
7 Lynn J. When does quality improvement count as research? Human subject protection and theories of knowledge. Quality and Safety in Health Care 2004;**13**(1):67-70.
8 Greenhalgh T, Russell J, Swinglehurst D. Narrative methods in quality improvement research. Quality & Safety in Health Care 2005;**14**(6):443-9 doi: 10.1136/qshc.2005.014712[published Online First: Epub Date].
9 Davidoff F, Batalden P, Stevens D, et al. Publication guidelines for improvement studies in health care: evolution of the SQUIRE Project. Annals of Internal Medicine 2008;**149**(9):670-6.
10 Verdú A, Maestre A, López P, et al. Clinical pathways as a healthcare tool: design, implementation and assessment of a clinical pathway for lower-extremity deep venous thrombosis. Quality and Safety in Health Care 2009;**18**(4):314-20.
11 May F, Simpson D, Hart L, et al. Experience with academic detailing services for quality improvement in primary care practice. Quality and Safety in Health Care 2009;**18**(3):225-31.
12 Fulop N, Protopsaltis G, King A, et al. Changing organisations: a study of the context and processes of mergers of health care providers in England. Social Science & Medicine 2005;**60**(1):119-30.
13 Rotter T, Kinsman L, James E, et al. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database of Systematic Reviews (Online) 2010;**3**(3) doi: 10.1002/14651858.CD006632.pub2.

14 O'Brien M, Rogers S, Jamtvedt G, et al. Educational outreach visits: effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviews (Online) 2007;**4**(4):1-62.
15 Haynes B. Can it work? Does it work? Is it worth it?: the testing of healthcare interventions is evolving. BMJ: British Medical Journal 1999;**319**(7211):652-63.
16 Franke RH, Kaul JD. The Hawthorne experiments: first statistical interpretation. American Sociological Review 1978;623-43.
17 Pirsig R. *Zen and the art of motorcycle maintenance: an enquiry into values*. New York: Bantam Books; 1984.

Capítulo 15 Aplicación de la práctica basada en la evidencia

¿Por qué los profesionales de la salud son lentos a la hora de adoptar la práctica basada en la evidencia?

El hecho de que los profesionales sanitarios no ejerzan según la mejor evidencia disponible no se puede atribuir por completo a la ignorancia o la testarudez. El pediatra Dr. Van Someren[1] ha descrito un ejemplo (que se ha convertido en histórico) que ilustra muchos de los obstáculos adicionales para aplicar la evidencia de la investigación en la práctica: la prevención del síndrome de dificultad respiratoria neonatal en los bebés prematuros.

En 1957, se descubrió que los bebés que nacen más de 6 semanas antes de la fecha prevista de parto pueden presentar graves dificultades respiratorias debido a la falta de una sustancia llamada surfactante, que disminuye la tensión superficial en los alvéolos pulmonares y reduce la resistencia a la expansión. En la década de 1960, las compañías farmacéuticas comenzaron la investigación para desarrollar un surfactante artificial que se pudiera administrar al lactante para evitar el desarrollo de este síndrome potencialmente mortal, pero no fue hasta mediados de la década de 1980 cuando se desarrolló un producto eficaz.

A finales de la década de 1980, se realizaron varios ensayos aleatorizados y un metaanálisis publicado en 1990 sugirió que los beneficios del surfactante artificial superaban en gran medida sus riesgos. En 1990 se inició un ensayo con 6.000 pacientes (OSIRIS), con la participación de casi todas las grandes unidades de cuidados intensivos neonatales de Reino Unido. El fabricante obtuvo una licencia para el producto en 1990 y en 1993 prácticamente todos los niños prematuros elegibles en el Reino Unido recibían surfactante artificial.

Una generación antes, también se había mostrado que otro tratamiento prevenía el síndrome de dificultad respiratoria neonatal: la administración del esteroide dexametasona a las madres con parto prematuro. La dexametasona actuaba acelerando la velocidad a la que el pulmón fetal alcanzaba la madurez. Su eficacia se había demostrado en animales de experimentación en 1969 y en ensayos clínicos en seres humanos, publicados en la prestigiosa revista *Pediatrics*, ya en 1972. Sin embargo, a pesar de un efecto beneficioso significativo confirmado en varios ensayos adicionales y de un metaanálisis publicado en 1990, la adopción de esta

tecnología fue increíblemente lenta. En 1995 se estimó que sólo el 12-18% de las madres elegibles recibían este tratamiento en EE.UU.[2].

La calidad de la evidencia y la magnitud del efecto eran similares para ambas intervenciones[3,4]. ¿Por qué los pediatras fueron mucho más rápidos que los obstetras a la hora de implementar una intervención que prevenía muertes evitables? El Dr. Van Someren[1] evaluó una serie de factores, que se enumeran en la tabla 15.1. El efecto del surfactante artificial es prácticamente inmediato y el médico que lo administra observa directamente la curación de un bebé con una enfermedad mortal sin tratamiento. El apoyo de la industria farmacéutica a un gran ensayo (y, probablemente, innecesario desde el punto de vista científico) garantizó que pocos pediatras que obtuvieron su título a principios de la década de 1990 ignorasen la introducción de la nueva tecnología.

En cambio, los esteroides, especialmente para las mujeres embarazadas, estaban pasados de moda y los pacientes los consideraban malos para la salud. En opinión de los médicos, la dexametasona era un tratamiento anticuado para varias enfermedades poco glamurosas, en especial el cáncer terminal, y el mecanismo científico de su efecto sobre los pulmones fetales no se entendía fácilmente. Lo más llamativo es que el obstetra pocas veces podía presenciar directamente el efecto de salvar la vida en un paciente concreto.

El ejemplo mencionado anteriormente no es un caso aislado. Algunas estrategias eficaces de asistencia sanitaria a menudo (aunque, por suerte, no siempre) tardan

Tabla 15.1 Factores que influyen en la aplicación de la evidencia para prevenir el síndrome de dificultad respiratoria neonatal (Dr. V. Van Someren, comunicación personal)

	Tratamiento con surfactante	*Tratamiento prenatal con esteroides*
Percepción del mecanismo	Corrige una enfermedad por deficiencia de surfactante	Efecto mal definido sobre el tejido pulmonar en desarrollo
Aparición del efecto	Minutos	Días
Impacto en el prescriptor	Observa directamente el efecto (tiene que estar de pie junto al ventilador)	Observa el efecto como una estadística en el informe anual
Percepción de los efectos secundarios	Percibidos como mínimos	Ansiedad del médico y los pacientes desproporcionada respecto al riesgo real
Conflicto entre dos pacientes	No (el paciente pediátrico se beneficiará directamente)	Sí (la paciente obstétrica no se beneficiará directamente)
Interés de la industria farmacéutica	Alto (producto patentado; enorme potencial de ingresos)	Bajo (producto fuera de patente; pequeños ingresos potenciales)
Tecnología de ensayos clínicos	«Nueva» (desarrolla a finales de la década de 1980)	«Antigua» (desarrollada a principios de la década de 1970)
Amplia participación de los médicos en los ensayos	Sí	No

años en popularizarse, incluso entre los expertos que deberían estar a la vanguardia de la práctica[5-8]. En las secciones restantes de este capítulo se expone cómo se puede reducir el tiempo desde que aparece la evidencia de investigación hasta que se realizan modificaciones reales en los resultados de salud. El lector debe tener en cuenta que no hay soluciones rápidas.

¿Cuánto sufrimiento evitable se debe a que no se aplica la evidencia?

La respuesta corta a esta pregunta es «mucho». Recientemente descubrí un artículo de Woolf y Johnson[9] en la revista *Annals of Family Medicine,* titulado: «The break-even point: when medical advances are less important than improving the fidelity with which they are delivered» (El punto de equilibrio: cuando los avances médicos son menos importantes que la mejora de la fidelidad con que se presentan). Su argumento era el siguiente: imagine una enfermedad que mata a 100.000 personas al año. Si se demuestra mediante la investigación que el fármaco X es eficaz para esta enfermedad, con una reducción de la mortalidad del 20%, podrá salvar 20.000 vidas anuales, pero si sólo el 50% de los pacientes elegibles reciben realmente el fármaco, el número de vidas salvadas se reduce a 10.000. Los autores alegan que, en muchos casos, añadiríamos más valor si aumentáramos nuestros esfuerzos para poner en práctica esta evidencia en lugar de llevar a cabo más investigaciones para desarrollar un fármaco diferente cuya eficacia sea mayor que la del fármaco X.

Para quien piense que estas cifras son especulativas, a continuación se presenta un ejemplo real citado del artículo de Woolf y Johnson, donde presentan la evidencia de un metaanálisis sobre los efectos de la aspirina en el ictus agudo y un estudio sobre las prácticas de prescripción en EE.UU.:

> *En una revisión sistemática de la Antithrombotic Trialists Collaboration se describió que el uso de aspirina en pacientes que habían sufrido previamente un ictus o un ataque isquémico transitorio reduce la incidencia de ictus no mortales recidivantes en un 23%. Es decir, en una población en la que 100.000 personas fuesen a sufrir accidentes cerebrovasculares, podrían evitarse 23.000 casos si todos los pacientes elegibles tomasen aspirina. Sin embargo, McGlynn y cols. describieron que el tratamiento antiagregante plaquetario sólo se usa en el 58% de los pacientes elegibles. Con esa cifra, sólo se pueden prevenir 13.340 ictus en la población hipotética, mientras que lograr un 100% de fidelidad a la hora de ofrecer la aspirina podría prevenir 23.000 ictus (es decir, 9.660 accidentes cerebrovasculares adicionales)*[9].

En resumen, la cantidad de sufrimiento evitable causado por no aplicar la evidencia se desconoce, pero podría calcularse utilizando el método indicado en el artículo de Woolf y Johnson. Es alentador que un porcentaje creciente (aunque aún pequeño) de los fondos de investigación ahora se asigne a aumentar la proporción de pacientes que se benefician de actuaciones cuya eficacia se conoce.

¿Cómo se puede influir en la conducta de los profesionales sanitarios para potenciar la práctica basada en la evidencia?

El grupo Cochrane Effective Practice and Organisation of Care (práctica y organización sanitaria eficaces, EPOC por su acrónimo en inglés, que se describe en el capítulo 10 y en la página de internet http://epoc.cochrane.org) ha realizado un trabajo exhaustivo de resumir la literatura acumulada a partir de ensayos de investigación sobre lo que es y lo que no es eficaz para cambiar la práctica profesional, tanto para promover innovaciones eficaces como para alentar a los profesionales a resistirse a las «innovaciones» que son ineficaces o perjudiciales. El grupo EPOC se ha interesado principalmente en revisar los ensayos de intervenciones destinadas a subsanar las posibles deficiencias en la secuencia de llevar la evidencia a la práctica.

Uno de los pocos mensajes inequívocos del trabajo del grupo EPOC es que simplemente *hablar* a la gente sobre la medicina basada en la evidencia (MBE) es siempre ineficaz para modificar la práctica. Hasta hace relativamente poco, la educación (por lo menos en relación con la formación de los médicos) era más o menos sinónimo de sesiones de tipo clase magistral que la mayoría de nosotros recordamos de la escuela y la universidad. El método de «calentar asiento» para la formación de posgrado (llenar auditorios con médicos o enfermeras para escuchar a un «experto» que imparte perlas de sabiduría) es relativamente barato y cómodo para los educadores, pero no suele lograr una modificación de la conducta sostenida en la práctica. De hecho, un estudio demostró que el número descrito de horas cursadas de FMC (formación médica continua) se correlacionaba *inversamente* con la competencia de los médicos[10].

Los lectores que estén interesados en la teoría en la que se sustenta la docencia de la MBE habrán observado que el «método instruccional» para promover el cambio de la conducta profesional respecto a la MBE se basa en la suposición errónea de que la gente se comporta de una manera particular *porque (y sólo porque) carece de conocimientos,* por lo que la transmisión de conocimientos modificará la conducta. La crítica corta y fidedigna de Marteau y cols.[10] muestra que este modelo no tiene ni coherencia teórica ni apoyo empírico. Estos autores concluyen que la información puede ser *necesaria* para el cambio de conducta profesional, pero pocas veces, o nunca, es *suficiente.* A continuación se citan teorías psicológicas en las que, en opinión de Marteau y su equipo, se podría basar el diseño de estrategias educativas más eficaces:

- *Aprendizaje conductual:* defiende la idea de que una conducta tiene más probabilidades de repetirse si se asocia con recompensas y menos si se asocia a un castigo.
- *Cognición social:* al planear una acción, las personas se preguntan: «¿vale la pena el esfuerzo?», «¿qué piensan otras personas sobre esto?» y «¿soy capaz de lograrlo?».
- *Modelos de las etapas de cambio:* en estos modelos se considera que todos los individuos están en algún punto de un continuo de preparación para el cambio, desde la ignorancia de que hay una necesidad de cambiar hasta la implementación sostenida de la conducta deseada.

Más recientemente, el equipo de Michie[11] amplió esta taxonomía simple con una mezcla heterogénea de otras teorías del cambio de conducta tomadas de la psicología cognitiva y el equipo de Eccles[12] (que incluye al gurú de las guías Jeremy Grimshaw) aplicó un conjunto similar de teorías psicológicas específicamente a la asimilación de la práctica basada en la evidencia por los médicos.

¿Qué tipo de enfoques educativos se ha demostrado que son eficaces en realidad para fomentar la práctica basada en la evidencia? A continuación se presenta un resumen de la literatura empírica, basado principalmente en cuatro revisiones sistemáticas de ensayos de intervención[13-16]:

(a) La enseñanza de la MBE como se realiza convencionalmente en los planes de estudio del grado de medicina mejora el conocimiento y las actitudes de los estudiantes respecto a la MBE, pero no se ha demostrado de manera convincente que su actuación tenga un impacto a la hora de tratar casos reales.
(b) En cuanto a los médicos titulados, la mayor parte de la formación en MBE basada en clases tiene poco o ningún impacto en sus conocimientos o habilidades de evaluación crítica. Esto puede deberse tanto a la formación como a que los exámenes no son obligatorios, o puede ser porque la propia formación es demasiado escasa, superficial, formulista, pasiva y también alejada de la práctica.
(c) Los métodos más sólidos desde el punto de vista educativo, como la enseñanza de MBE «integrada» (p. ej., durante los pases de planta o en el servicio de urgencias) o cursos cortos intensivos que utilizan métodos de aprendizaje muy interactivos, pueden producir cambios significativos en el conocimiento, las aptitudes y las conductas.
(d) Sin embargo, aún no se ha demostrado que estos cursos tengan ningún impacto directo sobre los resultados relevantes para los pacientes.

Green[17,18], que ha realizado uno de los estudios primarios más rigurosos sobre la formación en MBE jamás organizados, así como un estudio nacional de los programas y una revisión crítica, opina que la enseñanza de la MBE debería llevarse a cabo en un contexto real, es decir, en la consulta y a la cabecera del paciente. Este autor cita la teoría de la educación de adultos para respaldar el argumento de que la enseñanza de la MBE seguramente sería más eficaz si el alumno fuese capaz de relacionarla con problemas prácticos inmediatos y la utilizase para la toma de decisiones reales (en vez de hipotéticas). Él también ha realizado estudios cualitativos para confirmar que estas barreras prácticas del mundo real (falta de tiempo, evidencia inaccesible cuando es necesaria, cultura organizacional implacable, etc.) constituyen gran parte de la brecha entre la teoría y la práctica a la hora de aplicar la MBE[19]. Green sugiere que el camino que debe seguirse requiere realizar un trabajo aún mayor para asegurar que la evidencia está disponible y fácilmente accesible en el punto de atención, permitiendo que surjan preguntas clínicas y que se respondan en un contexto que optimice el aprendizaje activo.

En el capítulo 10 describí los principales resultados de la revisión sistemática de Grimshaw[20] de 2004 sobre la implementación de guías. La principal conclusión

de esta revisión fue que, a pesar de cientos de estudios que cuestan millones de dólares, ninguna intervención, ya sea educativa o de otro tipo, y de forma individual o combinada, *garantiza* que cambie la conducta de los profesionales hacia una dirección «basada en la evidencia».

En este punto es donde me separo ligeramente del enfoque del EPOC. Mientras que muchos miembros del EPOC siguen realizando ensayos (y revisiones de ensayos) para añadirlos a la base de investigación de si una u otra intervención (como folletos y otros materiales educativos impresos[21], la auditoría y la retroalimentación[22] o incentivos financieros[23,24]) es o no es eficaz para modificar la conducta clínica, mi opinión es que este esfuerzo es poco apropiado. No sólo porque no se hayan identificado aún soluciones mágicas, sino porque creo que *nunca se identificarán* y que debemos dejar de buscarlas.

Esto se debe a que la implementación de las mejores prácticas es muy compleja; implica múltiples influencias que actúan en diferentes direcciones[25] y depende de las personas. Un enfoque que tiene un efecto positivo en un estudio podría tener un efecto negativo en otro, por lo que el concepto de la «magnitud del efecto» de una intervención para cambiar la conducta clínica no sólo carece de sentido sino que es engañoso. Quien tenga hijos sabrá que una estrategia de crianza que haya funcionado bien para el primer hijo podría ser totalmente ineficaz para el segundo por razones que no son fáciles de explicar. Es algo que tiene que ver con la peculiaridad humana (el segundo hijo es un individuo diferente, con una personalidad distinta) y también con el hecho de que el contexto es sutilmente diferente de múltiples maneras, incluso en el «mismo» entorno familiar (el segundo hijo tiene un hermano mayor, padres más ocupados, juguetes de segunda mano, etc.). Lo mismo ocurre con las organizaciones, su personal y su práctica basada en la evidencia. Incluso el enfoque de investigación más refinado consistente en buscar «mediadores» y «moderadores» de la efectividad de las intervenciones particulares[12] sigue estando, en mi opinión, basado en la suposición errónea de que existe un efecto «mediador/moderador» constante de una variable contextual particular.

Insistamos un poco más en el factor humano. En una revisión sistemática de la difusión de innovaciones a nivel de organización en los servicios sanitarios, extraje esta conclusión acerca de los elementos humanos en la adopción de innovaciones:

> *Las personas no son receptores pasivos de las innovaciones. En su lugar (y en mayor o menor medida en los diferentes individuos) buscan las innovaciones, experimentan con ellas, las evalúan, las encuentran (o no) significado, desarrollan sentimientos (positivos o negativos) sobre ellas, las ponen en entredicho, se preocupan por ellas, se quejan de ellas, «trabajan en torno» a ellas, hablan con otros acerca de ellas, desarrollan conocimientos acerca de ellas, las modifican para adaptarlas a tareas particulares y tratan de mejorarlas o rediseñarlas*[25].

Éstos fueron los factores clave que mi equipo encontró que se asociaban con la disposición de una persona a adoptar innovaciones de asistencia sanitaria:

(a) *Antecedentes psicológicos generales:* varios rasgos de personalidad se asocian con la propensión a probar y utilizar innovaciones (p. ej., tolerancia a la ambigüedad, capacidad intelectual, motivación, valores y estilo de aprendizaje). En pocas palabras, algunas personas tienen unas costumbres más fijas que otras, por lo que necesitarán más estímulos y requerirán más tiempo para cambiar.

(b) *Antecedentes psicológicos específicos del contexto:* una persona que esté motivada y que sea capaz (en términos de valores, metas, habilidades específicas, etc.) de utilizar una innovación particular es más propensa a adoptarla. Además, si la innovación responde a una *necesidad identificada* en la persona señalada para adoptarla, será más probable que se adopte.

(c) *Significado:* el significado que la innovación presenta para la persona tiene una poderosa influencia en su decisión de adoptarla. El significado que se atribuye a una innovación no suele ser fijo, sino que puede negociarse y reformularse, por ejemplo, a través de conversaciones con otros profesionales u otras personas de la organización. Por ejemplo, en el ejemplo descrito en la sección «¿Por qué los profesionales de la salud son lentos a la hora de adoptar la práctica basada en la evidencia?», uno de los problemas probablemente fuese que el tratamiento con dexametasona fue visto inconscientemente por los médicos como un «tratamiento farmacológico paliativo anticuado, que se utilizaba en personas mayores». Para cambiar su práctica, tuvieron que ubicar este tratamiento en un nuevo esquema mental, como un «tratamiento preventivo actualizado, adecuado para las mujeres embarazadas».

(d) *Naturaleza de la decisión de adopción:* la decisión de un individuo perteneciente a una organización de adoptar una innovación particular pocas veces es independiente de otras decisiones. Puede ser supeditada (dependiente de una decisión tomada por otra persona de la organización), colectiva (el individuo tiene «voto», pero en última instancia, debe seguir la decisión de un grupo) o autorizada (a la persona se le dice si la adopta o no). Un buen ejemplo de la promoción de la práctica basada en la evidencia mediante una decisión de adopción autorizada es el desarrollo de formularios de medicamentos en los hospitales o consultas. Los fármacos que tienen una utilidad marginal o una relación coste-efectividad baja pueden ser retirados de la lista de medicamentos por los cuales el hospital está dispuesto a pagar. Sin embargo (como puede haber descubierto el lector si utiliza formularios de medicamentos impuestos), estas políticas también inhiben la práctica basada en la evidencia porque el innovador que vaya por delante de los demás debe esperar (a veces años) para que se tome una decisión del comité antes de implementar un nuevo estándar de práctica.

(e) *Preocupaciones y necesidades de información:* la gente se preocupa por cosas diferentes en etapas distintas de la adopción de una innovación. En un principio, necesitan *información general* (¿qué es la nueva práctica «basada en la evidencia», cuánto cuesta y cómo podría afectarme?); en las etapas iniciales de la adopción necesitan *información práctica* (¿cómo puedo hacer que funcione en la práctica?)

y, a medida que adquieren más confianza en la nueva práctica, necesitan *información de desarrollo y de adaptación* (¿puedo adaptar esta práctica un poco para que se ajuste a mis circunstancias? y, en caso afirmativo, ¿cómo debo hacerlo?).

Después de haber analizado la naturaleza de la idiosincrasia humana, otro factor importante que se debe tener en cuenta es la influencia que una persona puede tener sobre otra. Como Rogers[26] demostró por primera vez respecto a la adopción de innovaciones agrícolas por los agricultores de Iowa (que quizá tienen unas costumbres más arraigadas que los médicos), el contacto interpersonal es el método más poderoso de influencia. El tipo principal de influencia interpersonal respecto a la adopción de la práctica basada en la evidencia es el *líder de opinión.* Imitamos a dos tipos de personas: aquéllas a las que admiramos («líderes de opinión expertos») y aquéllas que consideramos iguales que nosotros (los «líderes de opinión compañeros»)[27].

Un líder de opinión que se oponga a una nueva práctica, o incluso uno que se muestre indiferente y no la respalde, tiene un gran potencial demoledor. Pero como se demostró en esta revisión sistemática de ensayos sobre la intervención de líderes de opinión, sólo porque un médico sea más propenso a cambiar su conducta de prescripción porque un líder de opinión respetado ya la haya cambiado, no se deduce necesariamente que por dirigir intervenciones educativas a los líderes de opinión (médicos designados por otros médicos como individuos a los que se consultaría o se imitaría) se produzca un cambio generalizado en la práctica de prescripción[28]. Esto se debe probablemente a que los líderes de opinión tienen su propia visión de las cosas y también a muchos otros factores que influyen en la práctica, aparte de una sola persona. En el mundo real, las denominadas «políticas de influencia social» pueden no tener influencia.

Otro modelo importante de influencia interpersonal, que la industria farmacéutica ha demostrado que es muy eficaz, es el contacto cara a cara entre los médicos y representantes de empresas farmacéuticas (comentados en el capítulo 6 y denominados visitadores médicos o representantes), cuya influencia sobre la conducta clínica puede ser tan determinante que se han apodado los «bombarderos invisibles» de la medicina. Como se muestra en el ejemplo de la sección «Diez preguntas que se deben plantear acerca de un artículo que describa una iniciativa de mejora de la calidad» del capítulo 14, esta táctica se ha materializado por las agencias de cambio no comerciales en lo que se denomina *visita académica:* el educador concierta una entrevista con el médico del mismo modo que los representantes farmacéuticos, pero en este caso el representante proporciona información objetiva, completa y comparativa sobre varios fármacos diferentes y anima al clínico a adoptar un enfoque crítico de la evidencia. Aunque en ensayos de investigación se han observado cambios espectaculares a corto plazo en la práctica, el ejemplo del capítulo anterior demuestra que en un contexto del mundo real los cambios positivos y constantes en la atención al paciente pueden ser difíciles de evidenciar[29]. Como siempre, la intervención no debe considerarse una panacea.

Una estrategia final que se debe tener en cuenta en relación con el apoyo de la implementación de la práctica basada en la evidencia es el uso de sistemas

computarizados de apoyo a la toma de decisiones que incorporan la evidencia de la investigación y a los que puede acceder el profesional atareado con tan sólo pulsar un botón. En la actualidad, se están desarrollando, probando y evaluando en ensayos clínicos aleatorizados decenas de estos sistemas, aunque relativamente pocos se usan de forma habitual. Se han realizado varias revisiones sistemáticas de estos sistemas, por ejemplo, la síntesis de Garg y cols.[30] de 100 estudios empíricos publicada en *JAMA* y la «revisión de revisiones» de Black y cols.[31] que englobó 13 revisiones sistemáticas previas sobre el apoyo a la toma de decisiones clínicas. Garg y cols. demostraron que alrededor de dos tercios de estos estudios mostraban una mejora del rendimiento clínico en el grupo de soporte de decisiones y los mejores resultados se obtuvieron en la posología de fármacos y la asistencia clínica activa (p. ej., tratamiento del asma) y los peores en el diagnóstico. Los sistemas que incluyeron un indicador espontáneo (en lugar de requerir que el clínico activase el sistema) y aquellos en los que el ensayo se llevó a cabo por las personas que desarrollaron la tecnología (en lugar de utilizar un producto comercial) fueron los más eficaces. En la revisión más reciente de Black y cols. se confirmaron ampliamente estos hallazgos. La mayoría de los estudios parecían demostrar mejoras significativas del rendimiento clínico (p. ej., seguir una guía, aplicación de cuidados preventivos como la vacunación o cribado del cáncer) mediante el apoyo a la toma de decisiones computarizada, pero el impacto sobre los resultados del paciente era mucho más variable. Estos últimos sólo se midieron en alrededor del 25% de los estudios y, cuando se midieron, por lo general mostraron un impacto modesto o nulo, excepto en el análisis de subgrupos a posteriori (que tiene una validez estadística cuestionable).

Hay que recordar lo que dije en un capítulo anterior (pág. 207) sobre la complejidad de la implementación de la MBE. Soy escéptica respecto a los estudios en los que se afirma que «el apoyo a la toma de decisiones computarizado es/no es eficaz» o «el apoyo a la toma de decisiones computarizado tiene un efecto de magnitud X». Dicho apoyo funciona para algunas personas en algunas circunstancias y nuestras energías de investigación deberían dirigirse actualmente a perfeccionar nuestras afirmaciones acerca de *qué tipo de* apoyo a la toma de decisiones computarizado, *para quién* y *en qué circunstancias*[32]. La resistencia de los médicos a las nuevas tecnologías es uno de mis intereses de investigación actuales, pero si escribiera todo lo que tengo que contar al respecto, este libro sería interminable. Desde aquí animo a los lectores que estén interesados a que consulten algunos los trabajos que publicaré en el plazo aproximado de un año.

¿Cómo es una organización «basada en la evidencia»?

Una de las preguntas que mi propio equipo abordó en nuestra revisión sistemática de la literatura sobre difusión de innovaciones a nivel organizacional fue: «¿cómo es una organización que promueve la adopción de innovaciones (basadas en la evidencia)?»[25]. Nosotros observamos que, en general, una organización asimilará un nuevo producto o práctica con más facilidad si es grande, madura (lleva mucho tiempo establecida), diferenciada funcionalmente (es decir, dividida en departamentos y

unidades semiautónomos), especializada (cuenta con una división bien desarrollada del trabajo, como servicios especializados), si tiene un excedente de recursos (dinero y personal) para dedicar a nuevos proyectos y si tiene las estructuras descentralizadas de toma de decisiones (los equipos pueden trabajar de forma autónoma). Aunque se han realizado docenas de estudios (y cinco metaanálisis) sobre el tamaño y la estructura de las organizaciones, todos estos determinantes representan menos del 15% de la variación en la capacidad de las organizaciones de apoyo a la innovación (y en muchos estudios, no explican nada de la variación en absoluto). Dicho de otro modo, no suele ser la estructura de la organización la que marca la diferencia fundamental en el apoyo a la MBE.

Un elemento más importante en nuestra revisión eran las dimensiones más difíciles de medir de la organización, sobre todo lo que los teóricos de las organizaciones denominan la *capacidad de absorción,* que se define como la capacidad de la organización para identificar, captar, interpretar, compartir, replantear y volver a codificar nuevos conocimientos, vincularlos con su propia base de conocimiento existente y darles un uso adecuado[33]. Entre los prerrequisitos para la capacidad de absorción se incluyen la base existente de conocimiento y habilidades de la organización (sobre todo su cantidad de conocimiento tácito de tipo práctico) y las tecnologías relacionadas preexistentes; una cultura de «organización de aprendizaje» (en la que se anima a la gente a aprender entre sí y compartir conocimientos), así como el liderazgo proactivo dirigido a permitir este intercambio de conocimientos[34].

En una extensa revisión de Dopson y cols.[35] sobre estudios cualitativos de alta calidad acerca de cómo se identifica, circula, se evalúa y se utiliza la evidencia de la investigación en las organizaciones sanitarias, se observó que antes de que pueda aplicarse plenamente en una organización, el conocimiento de MBE debe haberse adoptado y socializado, formando parte del *stock* de conocimiento que se desarrolla y se comparte socialmente con otras personas de la organización. Dicho de otro modo, el conocimiento depende para su circulación de las redes interpersonales (quién conoce a quién) y sólo se difundirá de manera eficiente a través de la organización si estas características sociales se tienen en cuenta y se superan las barreras.

Otra dimensión difícil de medir de una organización basada en la evidencia (es decir, una que es capaz de captar las mejores prácticas y aplicarlas de forma generalizada en la organización) es qué se conoce y un *contexto receptivo para el cambio.* Este constructo compuesto, desarrollado por Pettigrew y cols.[36] en relación con la aplicación de las mejores prácticas en la asistencia sanitaria, incorpora una serie de características organizacionales que se han asociado de forma independiente con su capacidad para adoptar nuevas ideas y afrontar la perspectiva de un cambio. Además de la capacidad de absorción de nuevos conocimientos (v. el texto anterior), los componentes del contexto receptivo son un liderazgo fuerte, una clara visión estratégica, unas buenas relaciones de gestión, personal visionario en puestos clave, un clima propicio para la experimentación y la asunción de riesgos, así como unos sistemas eficaces de captación de datos. El liderazgo puede ser especialmente crítico

para alentar a los miembros de la organización a romper con el pensamiento convergente y las rutinas que son la norma en organizaciones grandes y bien establecidas.

Otro artículo que merece la pena consultar es la revisión casi sistemática de Gustafson[37] sobre los determinantes de los proyectos de cambio exitosos en las organizaciones sanitarias. Algunos de los 18 ítems del modelo final de Gustafson son:

- Tensión para el cambio (el personal piensa que la práctica actual es subóptima y quiere que las cosas sean diferentes).
- Equilibrio de poder (el personal que apoya el cambio supera en número y tiene una posición más estratégica en la organización que el personal opuesto a él).
- Ventajas percibidas (todo el mundo entiende el cambio y cree que sus ventajas superan a los inconvenientes).
- Flexibilidad (la nueva práctica se puede adaptar para ajustarse a las necesidades y formas de trabajo locales).
- Tiempo y recursos (el cambio cuenta con la financiación adecuada y la gente tiene tiempo para trabajar en él).

Quien piense que esto es una receta que su organización no puede emplear en relación con la MBE, debería leer la sección siguiente (y si eso tampoco ayuda, debería pensar en cambiar de trabajo).

A quienes quieran saber más sobre artículos esenciales acerca de estudios de gestión y organización, les recomiendo el reciente resumen de la literatura del equipo de Ferlie[38] sobre temas como el conocimiento como recurso en las organizaciones (denominado en la jerga «perspectiva basada en los recursos de la organización») y los estudios críticos de gestión (un campo de investigación que plantea preguntas como «¿quién tiene el poder en esta organización?» y «¿a qué intereses beneficia este cambio?»), aplicados a la cuestión de si las organizaciones adoptan prácticas y políticas basadas en la evidencia y con qué rapidez las adoptan. Sus conclusiones son diversas, por lo que son difíciles de resumir, pero está claro que la comunidad de la MBE tiene mucho que aprender de nuestros colegas de disciplinas de gestión.

¿Cómo podemos ayudar a las organizaciones a desarrollar las estructuras, sistemas y valores adecuados para apoyar la práctica basada en la evidencia?

Aunque hay gran cantidad de evidencia sobre el tipo de organización que apoya la práctica basada en la evidencia, hay mucha menos sobre la eficacia de las intervenciones específicas para *cambiar* una organización de modo que esté más «basada en la evidencia» y queda fuera del alcance de este libro abordar este tema en profundidad. Gran parte de la literatura sobre el cambio organizacional corresponde a listas de comprobación prácticas o a un formato de tipo «diez consejos para el éxito». Las listas de comprobación y los consejos pueden ser de gran utilidad, pero esas listas tienden a crearme la necesidad de contar con algunos

modelos conceptuales coherentes en los que ubicar mis propias experiencias de la vida real.

La literatura de gestión ofrece no sólo una, sino varias docenas de distintos marcos conceptuales para observar el cambio. Esto deja confundidos a los lectores inexpertos acerca de por dónde empezar. Yo me propuse dar sentido a esta multiplicidad de teorías, para lo cual escribí una serie de seis artículos publicados hace unos años en el *British Journal of General Practice* titulada «Theories of change». En estos artículos, analicé seis modelos diferentes de cambio profesional y organizacional en relación con la práctica clínica eficaz[39-44]:

1. *Teoría del aprendizaje adulto:* el concepto de que los adultos aprenden a través de un ciclo de pensamiento y acción explica por qué la educación instruccional es tan constantemente ineficaz y por qué la experiencia práctica con la oportunidad de reflexionar y hablar con los colegas es la base fundamental tanto para el aprendizaje como para el cambio.
2. *Teoría psicoanalítica:* es el famoso concepto de Freud sobre el inconsciente, que influye (y a veces anula) nuestro yo consciente, racional. La resistencia de la gente al cambio a veces puede tener explicaciones emocionales poderosas y muy arraigadas.
3. *Teoría de las relaciones de grupo:* basada en los estudios realizados por especialistas de la clínica Tavistock de Londres sobre cómo actúan (o no actúan) los equipos en el entorno laboral. Las relaciones tanto dentro del equipo como entre el equipo y su entorno más amplio pueden actuar como barreras (o catalizadores) para el cambio.
4. *Teoría antropológica:* corresponde al concepto de que las organizaciones tienen culturas (es decir, formas de hacer las cosas y de pensar sobre los problemas) que suelen ser muy resistentes al cambio. Un cambio propuesto relativamente menor hacia la práctica basada en la evidencia (como requerir a los consultores que busquen evidencia de forma rutinaria en la base de datos Cochrane) puede ser en realidad muy amenazante para la cultura de la organización (en la cual, por ejemplo, la «opinión del consultor» ha tenido tradicionalmente un estatus casi sacerdotal).
5. *Teoría de la gestión clásica:* es el concepto de que «integrar» un cambio dentro de una organización requiere un plan sistemático para que ocurra. La visión para el cambio debe ser compartida por una masa crítica del personal y debe ir acompañada de modificaciones planeadas de las estructuras visibles de la organización, las funciones y responsabilidades de personas clave y de los sistemas de información y comunicación.
6. *Teoría de la complejidad:* es el concepto de que las grandes organizaciones (como el National Health Service de Reino Unido) dependen fundamentalmente de las relaciones dinámicas, evolutivas y locales, así como de los sistemas de comunicación entre las personas. El apoyo de las relaciones interpersonales clave y la mejora de la calidad y la oportunidad de la información disponible a nivel local suelen ser factores más decisivos a la hora de lograr un cambio sostenido que las directivas de tipo descendente o los programas generales nacionales o regionales.

Como ya he comentado, hay muchos modelos adicionales de cambio que pueden ser útiles para identificar y superar los obstáculos que impiden lograr una práctica basada en la evidencia. La lista mencionada previamente no pretende ser exhaustiva y, dada la complejidad de las organizaciones de asistencia sanitaria, ninguno de ellos ofrecerá una fórmula simple para que el cambio tenga éxito.

Yo añadiría un séptimo modelo teórico a la lista: el del cambio como un *movimiento social,* es decir, como una poderosa oleada de actividad que está ligada a la identidad de los individuos como parte del movimiento para el cambio[45]. Quien alguna vez haya estado en una manifestación de protesta, o se haya unido a una iniciativa vecinal para mejorar algún servicio local o de otro tipo, sabrá lo que se siente al ser parte de un movimiento social. Una vez estuve en un comité de alto nivel que trató de cerrar el poco usado servicio de urgencias de un hospital pequeño debido a que no había evidencia de que fuese eficaz o rentable, pero no conté con la aportación de la campaña «No toquéis nuestro hospital». De hecho, muchos cambios de éxito en la práctica clínica hacia la atención basada en la evidencia (p. ej., la abolición de la episiotomía sistemática en la atención obstétrica) se lograron principalmente a través de los grupos de presión de pacientes que funcionan a modo de «movimiento social».

Lo interesante de los movimientos sociales para el cambio es que, como subrayan Bate y cols.[45], aunque pueden lograr un cambio profundo y generalizado, no se puede planificar, controlar ni predecir su comportamiento del mismo modo que un modelo de gestión convencional. Aconsejo a los lectores que consulten el análisis sociológico de Pope[46] sobre el auge de la MBE como movimiento social.

Con independencia del enfoque teórico que se adopte para el cambio, convertir las teorías en práctica conllevará grandes dificultades. Una publicación de la National Association of Health Authorities and Trusts (NAHAT) de Reino Unido, titulada «Acting on the Evidence», hace hincapié en que la tarea de apoyar y capacitar a directivos y profesionales clínicos para utilizar la evidencia como parte de su toma diaria de decisiones es titánica y compleja[47]. En el apéndice 1 se presenta una lista de comprobación de acción para las organizaciones de asistencia sanitaria que trabajan hacia una cultura basada en la evidencia para la toma de decisiones clínicas y de formulación de políticas, adaptada de la publicación de la NAHAT.

En primer lugar, los principales responsables dentro de la organización, en particular los presidentes ejecutivos, consejeros y médicos más expertos, deben crear una cultura basada en la evidencia en la que se *espera* que la toma de decisiones se base en la mejor evidencia disponible. Se debe disponer de fuentes de información de alta calidad y actualizadas (como la biblioteca electrónica Cochrane y la base de datos Medline) en todas las oficinas, y se debe conceder tiempo al personal para acceder a ellas. Lo ideal es que los usuarios sólo tengan que usar un único punto de acceso para todas las fuentes disponibles. Se debe elaborar, difundir y utilizar información sobre la eficacia clínica y el coste-efectividad de tecnologías particulares. Las personas que recopilen y difundan esta información dentro de la organización deben conocer quién la utilizará y cómo se aplicará, y adaptar su presentación en consonancia. También deben establecer normas para

(y evaluar) la calidad de la evidencia que están ofreciendo. Las personas inscritas en la lista de correo interna de la organización para obtener información sobre efectividad requieren formación y apoyo para que puedan hacer el mejor uso de esta información.

Este consejo acertado de la NAHAT se basa (implícita o explícitamente) en el concepto de la *organización de aprendizaje.* Como Davies y Nutley[48] han señalado: «el aprendizaje es algo que logran los individuos, pero las "organizaciones de aprendizaje" pueden configurarse a sí mismas para maximizar, movilizar y mantener este potencial de aprendizaje». Basándose en el trabajo de Senge[49], ofrecen cinco características clave de una organización de aprendizaje:

1. Se alienta a las personas a superar las fronteras tradicionales, profesionales o departamentales (un enfoque que Senge denominó «pensamiento de sistemas abiertos»).
2. Las necesidades de aprendizaje personal de los individuos se identifican y se abordan sistemáticamente.
3. El aprendizaje se produce en cierta medida en los equipos, ya que es en gran parte a través de equipos cómo las organizaciones alcanzan sus objetivos.
4. Se hacen esfuerzos para cambiar la forma de conceptualizar los problemas, permitiendo nuevos enfoques creativos para viejos problemas.
5. Los médicos y gestores experimentados proporcionan liderazgo mediante una visión compartida con valores coherentes y una dirección estratégica clara, por lo que el personal colaborará voluntariamente para alcanzar un objetivo común.

Convertir una organización tradicional en una organización de aprendizaje es una tarea difícil, que a menudo implica un cambio importante en la cultura de la organización (las reglas no escritas, suposiciones y expectativas que constituyen «cómo se hacen aquí las cosas»). Aunque no es posible que un solo individuo modifique una organización, una persona que ocupe un cargo lo bastante alto como para escribir la descripción de las funciones de un nuevo miembro del personal, para decidir cómo se gasta el presupuesto de formación o para elegir quien interviene en un decisión clave puede comenzar a llevar a su organización en la dirección correcta (tabla 15.2).

Un principio básico en el desarrollo de una organización de aprendizaje es *invertir en las personas.* Además de un fuerte liderazgo en la cúspide de la organización, hay algunas funciones específicas a las que se debería apoyar en relación con la MBE[25]:

1. *Gestores de conocimiento:* se trata de personas experimentadas cuya función no es sólo obtener los sistemas de información, sino animar al resto del personal a utilizarlos. Toman las decisiones acerca de qué licencias de software se deben comprar para la organización y qué miembros del personal reciben permiso para el acceso a cada fuente de conocimiento. Cuando escribí la primera edición de este libro en 1995, una minoría de los hospitales tenía una regla según la cual el personal de enfermería no podía entrar en la biblioteca médica ni acceder a la conexión a internet. El papel del gestor del conocimiento es eliminar este tipo

Tabla 15.2 Diferencias clave entre una organización tradicional y una organización de aprendizaje

Característica	Organización tradicional	Organización de aprendizaje
Límites organizativos	Claramente delimitados	Permeables
Estructura de la organización	Prediseñada y fija	Evolutivos
Enfoque de los recursos humanos	Conjunto de habilidades mínimo para realizar el trabajo	Maximizar las habilidades para potenciar la creatividad y el aprendizaje
Enfoque de las actividades complejas	División en tareas segmentadas	Garantizar procesos integrados
Divisiones y departamentos	Grupos funcionales y jerárquicos	Redes abiertas, multifuncionales

Fuente: Senge[49]. Reproducida con autorización del Emerald Group Publishing Limited.

de normas sin sentido y asegurarse de que (en el caso de la MBE) todos quienes deban ejercerla tengan acceso a la base de conocimientos relevante, tiempo para acceder a ella y la formación adecuada.

2. *Trabajadores del conocimiento:* el trabajo de estas personas consiste en ayudar al resto a encontrar y aplicar el conocimiento. Un responsable del servicio de asistencia informática es un tipo de trabajador del conocimiento, al igual que un bibliotecario o un asistente de investigación. Para utilizar la jerga contemporánea, las herramientas de la MBE se deberían ofrecer como un producto aumentado, con una dotación de miembros del personal designado para proporcionar un apoyo flexible a los individuos como y cuando se solicite.
3. *Promotores:* la adopción de una nueva práctica por las personas de una organización o grupo profesional es más probable si las personas clave dentro de ese grupo están dispuestas a respaldar la innovación. El «respaldo» de una innovación basada en la evidencia podría incluir, por ejemplo, hablar con entusiasmo de ella, mostrar a la gente cómo usarla, añadirla a la agenda de los comités clave, conceder al personal tiempo para aprender sobre ella y probarla, y recompensar a las personas que la adopten. Aunque hay muy poca evidencia de investigación sobre lo que realmente hacen los promotores (o cuál es la forma más eficaz de impulsar un cambio basado en la evidencia), el principio es bastante simple: designar a personas concretas en todos los niveles de la organización para respaldarla.
4. *Ampliadores de fronteras (boundary spanners):* es más probable que una organización adopte un nuevo enfoque de la práctica si se pueden identificar individuos que tengan vínculos sociales significativos, tanto dentro como fuera de la organización, y que sean capaces y estén dispuestos a vincular a la organización con el mundo exterior en relación a esta práctica en particular. Tales individuos desempeñan un papel fundamental en la captación de las

ideas que se convertirán en innovaciones de la organización. Las organizaciones que tengan un miembro del personal que esté bien comunicado en un aspecto de la práctica basada en la evidencia deberían proponerse aprovechar sus conexiones y experiencia. Hay que enviar al personal fuera de la organización (a conferencias, visitas a organizaciones similares o para colaboraciones de mejora de calidad) y, al regreso, averiguar lo que han aprendido dedicando tiempo a escuchar sus historias e ideas.

Una herramienta específica que se debe tener en cuenta cuando se trabaja para lograr una «organización basada en la evidencia» es la idea de vías clínicas integradas, descrita como planes predefinidos de asistencia al paciente respecto a un tipo específico de diagnóstico (p. ej., sospecha de fractura de cadera) o de intervención (p. ej., reparación de hernia), con el objetivo de lograr un tratamiento más estructurado, coherente y eficaz[50]. He incluido un ejemplo de un intento de introducir una vía de este tipo en la sección «Diez preguntas que deben plantearse acerca de un artículo que describa una iniciativa de mejora de la calidad». Una buena vía clínica integra recomendaciones basadas en la evidencia con la realidad de los servicios locales, por lo general mediante una iniciativa multiprofesional que implica tanto a médicos como a gestores. La vía clínica no sólo indica qué intervención se recomienda en las diferentes etapas en la evolución de la enfermedad, sino también de quién es la responsabilidad de emprender la tarea y de realizar el seguimiento si no se lleva a cabo. Aunque hay muchas vías clínicas en circulación, el proceso de desarrollo de la vía suele implicar al personal de toda la organización en la misma medida que el producto finalizado a que se centre en una asistencia basada en la evidencia respecto a la enfermedad de interés. Cuando una organización sea refractaria a todo el concepto de la MBE, es posible que el proceso de desarrollar una vía clínica para una afección relativamente poco polémica logre una sorprendente cantidad de buena voluntad e implicación hacia el principio de la práctica basada en la evidencia, que se puede aprovechar para desarrollar la idea de forma más generalizada.

Por último, hay que señalar que el Health Service and Delivery Research Programme del National Institute for Health Research de Reino Unido (v. http://www.netscc.ac.uk/hsdr/) está financiando una amplia serie de estudios empíricos sobre el desarrollo, prestación y organización de servicios de salud, muchos de ellos de gran relevancia para la aplicación de las mejores prácticas a nivel organizacional. En la actualidad, hay más de 300 publicaciones de estudios de investigación sobre la implementación de la evidencia que se pueden descargar de forma gratuita.

Bibliografía

1 Van Someren V. *Changing clinical practice in the light of the evidence: two contrasting stories from perinatology. Getting research findings into practice*. London: BMJ Publications; 1994.

2 Gilstrap LC, Christensen R, Clewell WH, et al. Effect of corticosteroids for fetal maturation on perinatal outcomes. NIHconsensus development panel on the effect of corticosteroids

for fetal maturation on perinatal outcomes. JAMA: The Journal of the American Medical Association 1995;**273**(5):413-8.
3 Crowley PA. Antenatal corticosteroid therapy: a meta-analysis of the randomized trials, 1972 to 1994. American Journal of Obstetrics and Gynecology 1995;**173**(1):322-35.
4 Halliday H. Overviewof clinical trials comparingnatural andsynthetic surfactants. Neonatology 1995;**67**(Suppl. 1):32-47.
5 Booth-Clibborn N, Packer C, Stevens A. Health technology diffusion rates. International Journal of Technology Assessment in Health Care 2000;**16**(3):781-6.
6 Chauhan D, Mason A. Factors affecting the uptake of new medicines in secondary care–a literature review. Journal of Clinical Pharmacy and Therapeutics 2008;**33**(4):339-48.
7 Garjón FJ, Azparren A, Vergara I, et al. Adoption of new drugs by physicians: a survival analysis. BMC Health Services Research 2012;**12**(1):56.
8 Robert G, Greenhalgh T, MacFarlane F, et al. Organisational factors influencing technology adoption and assimilation in the NHS: a systematic literature review. Report for the National Institute for Health Research Service Delivery and Organisation programme, 2009, London.
9 Woolf SH, Johnson RE. The break-even point: when medical advances are less important than improving the fidelity with which they are delivered. The Annals of Family Medicine 2005;**3**(6):545-52.
10 Caulford PG, Lamb SB, Kaigas TB, et al. Physician incompetence: specific problems and predictors. Academic Medicine 1994;**69**(10):S16-8.
11 Michie S, Johnston M, Francis J, et al. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. Applied Psychology 2008;**57**(4):660-80.
12 Eccles M, Grimshaw J, Walker A, et al. Changing the behavior of healthcare professionals: the use of theory in promoting the uptake of research findings. Journal of Clinical Epidemiology 2005;**58**(2):107-12.
13 Horsley T, Hyde C, Santesso N, et al. Teaching critical appraisal skills in healthcare settings Cochrane Database of Systematic Reviews. The Cochrane Library 2011;(05): 2001;(3): CD001270.
14 Taylor R, Reeves B, Ewings P, et al. A systematic review of the effectiveness of critical appraisal skills training for clinicians. Medical Education 2000;**34**(2):120-5.
15 Coomarasamy A, Khan KS. What is the evidence that postgraduate teaching in evidence based medicine changes anything? A systematic review. BMJ: British Medical Journal 2004;**329**(7473):1017.
16 Norman GR, Shannon SI. Effectiveness of instruction in critical appraisal (evidence-based medicine) skills: a critical appraisal. Canadian Medical Association Journal 1998;**158**(2):177-81.
17 Green ML. Evidence-based medicine training in internal medicine residency programs. Journal of General Internal Medicine 2000;**15**(2):129-33.
18 Green ML. Evidence-based medicine training in graduate medical education: past, present and future. Journal of Evaluation in Clinical Practice 2000;**6**(2):121-38.
19 Green ML, Ruff TR. Why do residents fail to answer their clinical questions? A qualitative study of barriers to practicing evidence-based medicine. Academic Medicine 2005;**80**(2):176-82.
20 Grimshaw J, Thomas R, MacLennan G, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. Health Technology Assessment 2004;**8**:1-72.

21 Giguère A, Légaré F, Grimshaw J, et al. Printed educational materials: effects on professional practice and healthcare outcomes. Cochrane Database of Systematic Reviews 2012;**10**.
22 Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. Medical Care 2009;**47**(3):356-63.
23 Flodgren G, Eccles MP, Shepperd S, et al. An overview of reviews evaluating the effectiveness of financial incentives in changing healthcare professional behaviours and patient outcomes. Cochrane Database of Systematic Reviews (Online) 2011;**7**.
24 Scott A, Sivey P, Ait Ouakrim D, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database of Systematic Reviews (Online) 2011;**9**.
25 Greenhalgh T, Robert G, Macfarlane F, et al. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Quarterly 2004;**82**(4):581-629.
26 Rogers E. *Diffusion of innovations. 4th edition.* New York: Simon and Schuster; 2010.
27 Locock L, Dopson S, Chambers D, et al. Understanding the role of opinion leaders in improving clinical effectiveness. Social Science & Medicine 2001;**53**(6):745-57.
28 Flodgren G, Parmelli E, Doumit G, et al. Local opinion leaders: effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviews (Online) 2011;**8**.
29 Fischer MA, Avorn J. Academic detailing can play a key role in assessing and implementing comparative effectiveness research findings. Health Affairs 2012;**31**(10):2206-12.
30 Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. JAMA: The Journal of the American Medical Association 2005;**293**(10):1223-38.
31 Black AD, Car J, Pagliari C, et al. The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS Medicine 2011;**8**(1): e1000387.
32 Wong G, Greenhalgh T, Westhorp G, et al. RAMESES publication standards: realist syntheses. BMC Medicine 2013;**11**:21 doi: 10.1186/1741-7015-11-21[published Online First: Epub Date].
33 Zahra SA, George G. Absorptive capacity: a review, reconceptualization, and extension. Academy of Management Review 2002;**27**(2):185-203.
34 Ferlie E, Gabbay J, Fitzgerald L, et al. Evidence-basedmedicine and organisational change: an overview of some recent qualitative research, 2001.
35 Dopson S, FitzGerald L, Ferlie E, et al. No magic targets! Changing clinical practice to become more evidence based. Health Care Management Review 2010;**35**(1):2-12.
36 Pettigrew AM, Ferlie E, McKee L. *Shaping strategic change: making change in large organizations: the case of the National Health Service*. London: Sage, 1992.
37 Gustafson DH, Sainfort F, Eichler M, et al. Developing and testing a model to predict outcomes of organizational change. Health Services Research 2003;**38**(2):751-76.
38 Ferlie E, Crilly T, Jashapara A, et al. Knowledge mobilisation in healthcare: a critical review of health sector and generic management literature. Social Science & Medicine 2012;**74**(8):1297-304.
39 Greenhalgh T. Change and the team: group relations theory. British Journal of General Practice 2000;**50**:262-3.
40 Greenhalgh T. Change and the organisation 2: strategy. British Journal of General Practice 2000;**50**:424-5.
41 Greenhalgh T. Change and the organisation 1: culture and context. British Journal of General Practice 2000;**50**:340-1.

42 Greenhalgh T. Change and the individual 2: psychoanalytic theory. British Journal of General Practice 2000;**50**:164-5.
43 Greenhalgh T. Change and the individual 1: adult learning theory. British Journal of General Practice 2000;**50**:76-7.
44 Greenhalgh T. Change and complexity: the rich picture. British Journal of General Practice 2000;514-5.
45 Bate P, Robert G, Bevan H. The next phase of healthcare improvement: what can we learn from social movements? Quality and Safety in Health Care 2004;**13**(1):62-6.
46 Pope C. Resisting evidence: the study of evidence-based medicine as a contemporary social movement. Health 2003;**7**(3):267-82.
47 Appleby J, Walshe K, Ham C, et al. Acting on the evidence: a review of clinical effectiveness – sources of information, dissemination and implementation. NHS Confederation, Leeds, 1995.
48 Davies HT, Nutley SM. Developing learning organisations in the new NHS. BMJ: British Medical Journal 2000;**320**(7240):998.
49 Senge PM. The fifth discipline. Measuring Business Excellence 1997;**1**(3):46-51.
50 Rotter T, Kinsman L, James E, et al. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database of Systematic Reviews (Online) 2010;**3**(3).

Capítulo 16 Aplicación de la evidencia con los pacientes

La perspectiva del paciente

Este capítulo pretende demostrar precisamente que no existe *la* perspectiva del paciente. A veces, en nuestras vidas, a menudo con más frecuencia a medida que envejecemos, todos somos pacientes. Algunos de nosotros también somos profesionales sanitarios, pero cuando hay que tomar decisiones sobre *nuestra* salud, *nuestra* medicación, *nuestra* operación, o los efectos secundarios que *nosotros* podemos o no podemos experimentar con un tratamiento particular, contemplamos esa decisión de un modo diferente a cuando tomamos el mismo tipo de decisión en nuestro papel profesional.

Como ya sabrá quien haya leído los capítulos anteriores de este libro, la medicina basada en la evidencia (MBE) consiste principalmente en el uso de algún tipo de media poblacional (odds ratio, número necesario a tratar, estimación del tamaño del efecto medio, etc.) para documentar las decisiones. Sin embargo, muy pocas personas se comportan exactamente igual que el promedio de puntos de la gráfica: algunos serán más propensos a obtener los beneficios y otros lo serán a sufrir los perjuicios de una intervención en especial. Pocas personas valorarán un resultado particular en la misma medida que la media del grupo respecto a, por ejemplo, una pregunta de apuesta estándar (v. sección «Medición de los costes y beneficios de las intervenciones sanitarias» del capítulo 11).

La experiencia individual única de estar enfermo (o, en realidad, de estar «en riesgo» o de ser clasificado como tal) se puede expresar en términos narrativos: es decir, se puede contar una historia al respecto. Y la historia de cada uno es diferente. El «mismo» conjunto de síntomas o fragmento de información tendrá muchos significados diferentes dependiendo de quién los experimenta y de qué más está pasando en sus vidas. El ejercicio de realizar la historia clínica de un paciente es un intento de «domesticar» ese conjunto idiosincrásico e individual de experiencias personales y ponerlo en un formato más o menos estándar para hacerlo coincidir con los protocolos de evaluación, tratamiento y prevención de la enfermedad. De hecho, el primer profesor de medicina general de Inglaterra, Marshall Marinker[1], una vez afirmó que el papel de la medicina es distinguir

el mensaje claro de la enfermedad desde el ruido de interferencia del paciente como persona.

Como he escrito en otra parte, la perspectiva de la MBE sobre la *patología* y la perspectiva única del paciente sobre su *enfermedad* («medicina basada en la narrativa», si se quiere) no son, en absoluto, incompatibles[2].

Merece la pena volver a la definición original de la MBE propuesta por Sackett y cols. Esta definición se reproduce en su totalidad, aunque sólo suele citarse la primera frase:

> *La medicina basada en la evidencia es el uso consciente, explícito y sensato de la mejor evidencia actual en la toma de decisiones sobre la asistencia de los pacientes concretos. La práctica de la medicina basada en la evidencia significa integrar la experiencia clínica individual con la mejor evidencia clínica externa disponible procedente de la investigación sistemática. Con el término «experiencia clínica individual» nos referimos a la competencia y el criterio que cada médico adquiere mediante la experiencia y la práctica clínicas. Una mayor experiencia se refleja en muchos aspectos, pero especialmente en el diagnóstico más eficaz y eficiente, y en la identificación más reflexiva y el uso compasivo de las circunstancias, derechos y preferencias de los pacientes concretos en la toma de decisiones clínicas sobre su asistencia. La «mejor evidencia clínica externa disponible» es la investigación clínicamente relevante, a menudo procedente de las ciencias básicas de la medicina, pero sobre todo de la investigación clínica centrada en pacientes sobre la exactitud y la precisión de las pruebas de diagnóstico (incluida la exploración física), la potencia de los marcadores pronósticos y la eficacia y seguridad de las pautas terapéuticas, de rehabilitación y preventivas* (pág. 71).

Por lo tanto, aunque a veces se afirma erróneamente que los protagonistas originales de la MBE dejaron al pobre paciente fuera del guión, en realidad fueron muy cuidadosos para presentar la MBE como supeditada a la elección del paciente (y, de paso, como dependientes del criterio clínico). El «mejor» tratamiento no es necesariamente el que haya demostrado ser más eficaz en ensayos controlados aleatorizados, sino el que se adapte a un conjunto particular de circunstancias concretas y coincida con las preferencias y prioridades del paciente.

Los médicos a veces plantean el enfoque «basado en la evidencia» de un modo estereotipado, por ejemplo, al pensar que todos los pacientes con un accidente isquémico transitorio deben tomar warfarina ya que éste es el tratamiento preventivo más eficaz, con independencia de si los pacientes dicen que no quieren tomar comprimidos, no pueden tolerar los efectos secundarios o creen que no merece la pena la molestia de realizarse un análisis de sangre cada semana para comprobar su coagulación. Una de mis familiares era reacia a tomar warfarina, por ejemplo, porque le habían aconsejado que dejase de comer pomelo (un alimento que llevaba tomando en el desayuno desde hacía más de 60 años, pero que contiene productos

químicos que pueden interactuar con la warfarina). Me gustó que su médico de cabecera la invitase a comentar los pros y los contras de las diferentes opciones de tratamiento para que pudiese realizar una elección informada.

Casi toda la investigación en la tradición de la MBE entre 1990 y 2010 se centró en el componente epidemiológico y trató de establecer una base de evidencia de ensayos controlados aleatorizados y otros diseños de investigación con «solidez metodológica». Después, surgió una tradición de «elección del paciente basada en la evidencia», en la cual se formalizó y se estudió sistemáticamente el derecho del paciente a elegir la opción más adecuada y aceptable para él[3]. El tercer componente de la MBE al que se hace referencia en la cita (el criterio clínico individual) no ha sido teorizado ampliamente por los estudiosos de la tradición de la MBE, aunque yo he escrito acerca de él[4].

PROM

Antes de describir cómo se puede involucrar a los pacientes en la individualización de las decisiones de la MBE, quiero introducir un enfoque relativamente nuevo para seleccionar las medidas de resultado usadas en los ensayos clínicos: medidas de resultado referidas por los pacientes (PROM, *patient reported outcome measures*). Una posible definición es la siguiente:

> *Las PROM son las herramientas que utilizamos para obtener información desde la perspectiva del paciente sobre cómo se percibe que los aspectos de su salud y el impacto de la enfermedad y su tratamiento están influyendo en su estilo de vida y, posteriormente, en su calidad de vida (CDV). Por lo general son cuestionarios autocumplimentados, que pueden ser cumplimentados por un paciente o individuo sobre sí mismo, o por otras personas en su nombre*[5].

Con el término «medida de resultado» me refiero al aspecto de salud o enfermedad que los investigadores deciden medir para demostrar, por ejemplo, si un tratamiento ha sido eficaz. El fallecimiento es una medida de resultado, al igual que la presión arterial, la oportunidad de salir del hospital con un bebé vivo cuando se ingresa en un hospital para un parto o la capacidad de subir las escaleras y hacerse una taza de té uno mismo. Hay muchas más, pero la esencia es que en cualquier estudio los investigadores tienen que definir en qué están tratando de influir.

Las PROM no son medidas individualizadas. Al contrario, siguen siendo un tipo de media poblacional, pero a diferencia de la mayoría de las medidas de resultado, son un promedio de lo que más importa a los pacientes, en lugar de un promedio de lo que los investigadores o los médicos consideraron que debían medir. La manera de desarrollar una PROM es llevar a cabo una extensa fase de la investigación cualitativa (v. cap. 12) con una muestra representativa de personas que tengan la enfermedad en la cual estamos interesados, analizar los datos cualitativos y luego usarlos para diseñar un instrumento de estudio («cuestionario», v. cap. 13) que recoja todas las características clave de lo que preocupa a los pacientes[6,7].

Las PROM fueron popularizadas inicialmente (creo) por un equipo de Oxford dirigido por Ray Fitzpatrick, quien utilizó el concepto para desarrollar medidas dirigidas a evaluar el éxito de la artroplastia total de cadera y rodilla[8]. Ahora se usan de forma bastante rutinaria en muchos temas clínicos en el campo más amplio de «investigación de resultados»[9,10]; además, una monografía reciente realizada por el Kings Fund de Reino Unido recomienda su uso rutinario en la toma de decisiones del National Health Service[11]. Justo cuando esta edición se envió a imprenta, el *Journal of the American Medical Association* publicó un conjunto de normas para las PROM[12].

Toma de decisiones compartida

Por importante que sean las PROM, sólo indican qué es lo que más valoran los pacientes, en promedio, no lo que el paciente que tenemos frente a nosotros en un momento dado valora más. Para averiguarlo, como dije en el capítulo 1, habría que preguntar al paciente. En la actualidad, hay una ciencia y una metodología para «preguntar al paciente»[3,13].

La ciencia de la toma de decisiones compartida comenzó a finales de la década de 1990 como un interés especial de algunos médicos generales con un espíritu docente y entusiasta, sobre todo Elwyn y Edwards[14]. La idea se basa en el concepto del paciente como un selector racional, capaz y dispuesto (quizás con apoyo) a participar en la deliberación sobre las opciones y a realizar una elección informada.

Una de las dificultades es mantener la ecuanimidad, es decir, abstenernos de realizar lo que creemos que debería ser la forma de proceder y presentar las diferentes opciones presentando objetivamente los pros y los contras para que los pacientes puedan tomar sus propias decisiones[15]. En el cuadro 16.1 se enumeran las competencias que los médicos necesitan para poner en práctica la toma de decisiones compartida con sus pacientes[16].

Cuadro 16.1 Competencias para la toma de decisiones compartida (v. referencia 14)

Definir el problema: especificación clara del problema que requiere una decisión.
Mostrar ecuanimidad: los profesionales pueden no tener una preferencia clara sobre qué opción de tratamiento es la mejor en un contexto determinado.
Presentar las opciones: una o más opciones de tratamiento y la opción de no realizar ningún tratamiento en su caso.
Proporcionar información en el formato preferido: identificar las preferencias de los pacientes para que puedan ser de utilidad en el proceso de toma de decisiones.
Comprobar la comprensión: de la gama de opciones y de la información proporcionada sobre ellas.

Explorar ideas, preocupaciones y expectativas: sobre la enfermedad clínica, las opciones posibles de tratamiento y los resultados.
Comprobar el papel que se prefiera adoptar: que los pacientes acepten el proceso e identificar el papel que prefieren adoptar en la toma de decisiones.
Toma de decisiones: involucrar al paciente en la medida en que desee participar.
Aplazamiento si es necesario: revisar las necesidades de tratamiento y las preferencias después de un período para examinarlas más a fondo, incluso con amigos o miembros de la familia, si el paciente lo requiere.
Revisar lo acordado: período de tiempo especificado para revisar la decisión.

Los diversos instrumentos y herramientas para ayudar en la toma de decisiones compartida han evolucionado a lo largo de los años. Como mínimo, una ayuda para la toma de decisiones podría hacer que la información bastante árida de la MBE fuese más accesible para una persona no experta, por ejemplo, presentando los datos numéricos en forma de diagramas e imágenes[17]. En el ejemplo que se

Figura 16.1 Ejemplo de una ayuda en la toma de decisiones: elección de la estatina en un paciente diabético con un 20% de riesgo de infarto de miocardio. Fuente: reproducida de la referencia bibliográfica 18.

muestra en la figura 16.1, se utilizan colores e iconos simples para transmitir estimaciones cuantitativas del riesgo[18]. Las formas de medir el grado en que los pacientes han estado involucrados en una decisión también han evolucionado[19].

Coulter y Collins[20] han elaborado una excelente guía, *Making Shared Decision-Making a Reality*, que establece las características de una ayuda realmente buena para la toma de decisiones (cuadro 16.2).

Cada vez es más frecuente que se disponga de ayudas para la toma de decisiones en internet, que permiten al paciente hacer clic a través de diferentes pasos en el algoritmo de toma de decisiones (con o sin el apoyo de un profesional sanitario). En mi opinión, la mejor manera de acostumbrarse a las herramientas para la toma de decisiones compartidas es consultar varias de ellas y, si es posible, utilizarlas en la práctica. El National Health Service de Reino Unido tiene una página de internet con enlaces a herramientas para compartir decisiones, desde la reparación de un aneurisma de la aorta abdominal hasta la prevención del ictus en la fibrilación auricular: v. http://sdm.rightcare.nhs.uk/pda/. Se puede consultar un conjunto similar (y más completo) de herramientas para la toma de decisiones en la página canadiense de internet http://decisionaid.ohri.ca/AZinvent.php.

Tablas de opciones

Los estudios que utilizan la escala «OPTION» sugieren que la participación del paciente en la toma de decisiones basada en la evidencia no siempre es tan idónea como a los idealistas les gustaría que fuera[19]. En la actualidad, la mayoría de los

Cuadro 16.2 Características de una buena ayuda para la toma de decisiones (de la referencia 17)

Las ayudas para la toma de decisiones son diferentes de los materiales tradicionales de información al paciente porque no le dicen a la gente lo que debe hacer. En su lugar, exponen los hechos y ayudan a las personas a deliberar sobre las opciones. Por lo general contienen:

- Una descripción de la enfermedad y de los síntomas.
- El pronóstico probable con y sin tratamiento.
- Las opciones de tratamiento y de apoyo para el automanejo, así como las probabilidades de los resultados.
- Lo que se conoce a partir de la evidencia y lo que se ignora (incertidumbres).
- Ilustraciones para ayudar a las personas a entender qué se sentiría si se experimentasen algunos de los efectos secundarios o complicaciones más frecuentes de las opciones terapéuticas (a menudo utilizando entrevistas con pacientes).
- Un medio para ayudar a las personas a clarificar sus preferencias.
- Referencias y fuentes de información adicional.
- Credenciales, fuente de financiación y declaraciones de conflicto de intereses de los autores.

profesionales de la salud están (supuestamente) dispuestos a compartir las decisiones con los pacientes en principio, pero la investigación cualitativa y basada en cuestionarios ha demostrado que perciben una serie de barreras para hacerlo en la práctica, incluidas las limitaciones de tiempo y la falta de aplicabilidad del modelo de ayuda a la toma de decisiones a las características específicas de un paciente en particular[21]. Es relativamente infrecuente que los médicos remitan a los pacientes a páginas de internet de apoyo a la toma de decisiones, en parte porque piensan que ya están compartiendo las decisiones en la charla habitual de la consulta y en parte porque creen que los pacientes no desean participar de esta manera[22].

La realidad de una consulta típica de medicina general, por ejemplo, está muy alejada de la realidad objetiva de un algoritmo formal de toma de decisiones. Cuando un paciente acude con síntomas sugestivos de ciática, por ejemplo, el médico tiene 10 minutos para dedicarle. Por lo general, explorará al paciente, solicitará algunas pruebas y luego tendrá una conversación más bien difusa sobre cómo (por una parte) los síntomas del paciente podrían resolverse con fisioterapia, pero (por otra parte) querría solicitar una consulta con un especialista, ya que algunos casos necesitarán una operación. El paciente suele expresar una preferencia vaga por un tratamiento conservador o intervencionista y el médico (respetando los puntos de vista «facultados») aceptará la preferencia del paciente.

Si el médico está comprometido con la toma de decisiones compartida basada en la evidencia, puede intentar usar un enfoque más estructurado para la toma de decisiones compartida como se indica en la sección «Toma de decisiones compartida», por ejemplo, consultando un algoritmo en línea o utilizando gráficos circulares u hojas de cálculo preprogramadas para obtener puntuaciones numéricas de en qué medida el paciente puntúa procedimientos y resultados específicos respecto a otros. Sin embargo, con mucha frecuencia este tipo de herramientas se habrán probado una o dos veces y luego se habrán abandonado al considerarlas tecnocráticas, laboriosas, demasiado cuantitativas y extrañamente desconectadas de los relatos personales y específicos sobre la enfermedad que abundan en la consulta.

La buena noticia es que nuestros colegas que trabajan en el ámbito de la toma de decisiones compartida han reconocido recientemente que lo perfecto puede ser enemigo de lo bueno. La mayoría de las discusiones acerca de las opciones de tratamiento en la práctica clínica no necesitan un análisis exhaustivo de probabilidades, riesgos y puntuaciones de preferencia (e incluso pueden quedar desbaratadas por él). Lo que la mayoría de la gente quiere es una lista breve, pero equilibrada, de las opciones, que exponga los costes y beneficios de cada una, incluida una respuesta a la pregunta de qué pasaría si tomase esa decisión.

Se puede consultar una de estas tablas en la página http://www.optiongrid.org. Se trata de una iniciativa de colaboración entre pacientes, médicos y académicos[23]. Cada tabla de opciones ocupa una página y abarca un solo tema (hasta ahora se han realizado las correspondientes a la ciática, nefropatía crónica, cáncer de mama, amigdalitis y alrededor de una docena más). La tabla enumera las diferentes opciones en las columnas, mientras que en cada fila se responde a una pregunta

diferente (como «¿en qué consiste el tratamiento?», «¿cuándo me sentiré mejor?» y «¿cómo afectará este tratamiento a mi capacidad de trabajar?»). Un ejemplo se muestra en la figura 16.2.

Las tablas de opciones se desarrollan de forma similar a las PROM, pero suelen centrase más en la participación del equipo clínico multidisciplinario, como en este ejemplo de una tabla de opciones para el tratamiento del cáncer de la cabeza y el cuello[24]. La característica distintiva del enfoque de tabla de opciones es que promueve y apoya lo que se ha denominado *charla sobre las opciones,* es decir, las conversaciones y deliberaciones en torno a las diferentes opciones[25]. Las tablas tienen, en efecto, un diseño analógico en lugar de digital.

Ciática debida a una hernia de disco
Esta tabla está diseñada para ayudar al paciente y al médico a seleccionar la opción terapéutica que sea mejor para el paciente. Está indicada en personas diagnosticadas de hernia de disco que han tenido ciática durante al menos 6 semanas y no debe usarse en personas con problemas intestinales y urinarios debidos a compresión nerviosa por el disco. Se debe preguntar al profesional sanitario si existen otras opciones terapéuticas disponibles.

Preguntas frecuentes	Tratamiento sin inyecciones o cirugía	Inyecciones (esteroides epidurales)	Cirugía
¿En qué consiste el tratamiento?	Tomar analgésicos que reducen la inflamación alrededor del nervio e intentar realizar la máxima actividad posible. La fisioterapia también puede ayudar	Se utiliza una aguja para inyectar un anestésico local y un esteroide en el punto de compresión del nervio cerca de la columna vertebral. La inyección se suele realizar en una clínica especial y requiere unos 20 minutos	El disco herniado que comprime el nervio se extirpa durante una operación de la espalda. La operación dura unas 2 horas. La mayoría de las personas están en el hospital 1-2 noches, pero algunos se van a casa el día de la cirugía
¿Cuándo empezará a sentirse la mejoría?	6 semanas después del diagnóstico, alrededor del 20% de las personas dicen que están muy o algo satisfechas con sus síntomas	La mayoría de las personas que obtienen alivio se sienten mejor en la primera semana de la inyección aproximadamente	6 semanas después de la cirugía, alrededor del 60% de las personas dicen que están muy o algo satisfechas con sus síntomas
¿Qué tratamiento proporciona los mejores resultados a largo plazo?	1 año después del diagnóstico, alrededor del 45% de las personas que se tratan sin cirugía ni inyecciones dicen que están muy o algo satisfechas con sus síntomas	Es difícil decir: algunos estudios han demostrado beneficios de las inyecciones de esteroides, pero otros no	1 año después de la cirugía, alrededor del 70% de las personas dicen que están muy o algo satisfechas con sus síntomas
¿Cuáles son los principales riesgos/ efectos secundarios asociados con este tratamiento?	Todos los fármacos tienen algunos efectos secundarios. Realizar actividad es improbable que dificulte el tratamiento de la ciática en el futuro	Menos del 1% de las personas presentan complicaciones que podrían consistir en hemorragia, cefalea e infección	Los principales riesgos asociados con la cirugía son la infección (2%), la trombosis (1%) y la lesión de los nervios (menos del 1%)
¿Cómo influirá este tratamiento en mi capacidad de trabajar?	Se deberían reanudar las actividades diarias y el trabajo en cuanto sea posible	La mayoría de las personas reanudan las actividades diarias y el trabajo el día siguiente a la inyección	La mayoría de las personas están de baja 6-8 semanas después de esta cirugía
¿Será necesario otro tratamiento?	Hay que mantener la actividad. Es posible derivar al paciente para iniciar un programa de fisioterapia	Se deberían tomar analgésicos a demanda y mantener la actividad. La inyección se puede repetir en el futuro, por lo general, no más de 2-3 veces en total	La mayoría de las personas realizan fisioterapia después de la cirugía y toman analgésicos para controlar e dolor postoperatorio. En los años siguientes a la intervención, un pequeño número de pacientes requerirán más cirugía (alrededor del 5% en el primer año)

Figura 16.2 Ejemplo de una tabla de opciones. Fuente: http://www.optiongrid.org/optiongrids.php. Reproducida con autorización de Glyn Elwyn.

La razón por la que considero que este enfoque es un avance respecto a los enfoques más algorítmicos para la toma de decisiones compartida descritos en la sección «La perspectiva del paciente» es que la información en una tabla de opciones se presenta en un formato que permite la reflexión y el diálogo. La tabla se puede imprimir o se puede dar al paciente la dirección de internet e invitarle a que la visite y considere las opciones antes de regresar para una nueva consulta. Y a diferencia de la generación anterior de herramientas para la toma de decisiones compartida, ni el paciente ni el médico tienen que ser unos expertos de la informática para utilizarlas.

Ensayos *n* de 1 y otros enfoques individualizados

El último enfoque de la implicación de los pacientes que quiero presentar en este capítulo es el ensayo *n* de 1. Se trata de un diseño muy simple en el que cada participante recibe, en un orden asignado al azar, tanto la intervención como el tratamiento de control.

Esto probablemente se explica mejor con un ejemplo. Ya en 1994, algunos médicos generales australianos querían abordar la cuestión clínica de qué analgésico utilizar en la artrosis[26]. En su opinión, algunos pacientes tenían buenos resultados con paracetamol (que causa relativamente pocos efectos secundarios), mientras que otros no respondían tan bien al paracetamol, pero presentaban un gran alivio con un antiinflamatorio no esteroideo (AINE). En el contexto clínico normal, se podría probar con paracetamol primero y pasar al AINE si el paciente no respondía. Sin embargo, ¿qué sucedería si supusiésemos que habría un intenso efecto placebo? Es posible que el paciente tuviese una confianza limitada en el paracetamol ya que es un fármaco muy habitual, mientras que puede que prefiriese inconscientemente un AINE en un envase atractivo.

La idea del ensayo *n* de 1 es que todos los tratamientos se anonimizan, se preparan en formulaciones y envases idénticos, y simplemente se etiquetan A, B, etcétera. Los participantes no saben qué fármaco están tomando, por lo que su respuesta no está influenciada por si «creen» en el tratamiento. Para aumentar el rigor científico, los fármacos pueden tomarse de forma secuencial como ABAB o AABB, sin períodos de reposo farmacológico intermedios.

El ensayo *n* de 1 de March y cols. sobre el paracetamol frente a los AINE confirmó la sospecha clínica de que algunos pacientes mejoraron notablemente con el AINE, pero muchos también mejoraron igualmente con paracetamol. Es importante destacar que, a diferencia de un ensayo aleatorizado estándar, el diseño *n* de 1 permitió a los investigadores identificar qué pacientes estaban en cada categoría. Sin embargo, la tasa de abandono del ensayo fue alta, en parte porque cuando los participantes encontraban un fármaco eficaz, sólo querían seguir tomándolo en lugar de pasar a la alternativa.

Con todo, a pesar de su elegancia conceptual y de una promesa lejana de su relación con el paradigma de la «medicina personalizada» (donde en cada paciente se personalizarán las pruebas y las opciones terapéuticas en función

de su genoma, fisioma, microbioma, etc., específicos), el ensayo *n* de 1 no ha arraigado ampliamente en la investigación y la práctica clínica. Un artículo de revisión de Lillie y cols.[27] sugiere las causas. Dichos ensayos son muy laboriosos, pues requieren un alto grado de personalización individual y grandes cantidades de datos para cada participante. Los períodos de «reposo farmacológico» plantean problemas prácticos y éticos (¿hay que soportar la artritis sin analgésicos durante varias semanas para servir a la ciencia?). La combinación de los resultados de los diferentes participantes plantea dificultades estadísticas. Además, la (conceptualmente simple) ciencia de los ensayos *n* de 1 ha comenzado a confundirse con la ciencia mucho más compleja e incierta de la medicina personalizada.

En resumen, el ensayo *n* de 1 es un diseño útil (y sobre el cual puede preguntarse en los exámenes), pero no es la panacea que en su momento se había pretendido.

Moore y cols.[28] han propuesto recientemente un enfoque alternativo (y en parte no evaluado) a la individualización de los regímenes terapéuticos en relación con el alivio del dolor. Su argumento básico es que debería «esperarse el fracaso» (dado que el número necesario a tratar para muchas intervenciones es superior a 2, estadísticamente hablando es más probable que cualquier individuo *no* se beneficie a que obtenga beneficio), pero «perseguir el éxito» (porque la «media» de cualquier respuesta a una intervención enmascara un subgrupo de respondedores que mejorarán mucho con dicha intervención). Estos autores proponen un proceso de prueba y error guiado, probando sistemáticamente una intervención seguida de otra, hasta que se identifique la que funciona eficazmente en *cada* paciente. Tal vez éste sea el ensayo *n* de 1 sin preocuparse del factor placebo o del hecho de que tal vez se requiera probar media docena de opciones antes de encontrar la mejor en cada caso.

Bibliografía

1 Marinker M. The chameleon, the Judas goat, and the cuckoo. The Journal of the Royal College of General Practitioners 1978;**28**(189):199-206.

2 Greenhalgh T. Narrative basedmedicine: narrative based medicine in an evidence based world. BMJ: British Medical Journal 1999;**318**(7179):323.

3 Edwards A, Elwyn G. *Shared decision-making in health care: achieving evidencebased patient choice*. New York: Oxford University Press, 2009.

4 Greenhalgh T. Uncertainty and clinical method. Clinical Uncertainty in Primary Care: Springer 2013;23-45.

5 Meadows KA. Patient-reported outcome measures: an overview. British Journal of Community Nursing 2011;**16**(3):146-51.

6 Garratt A, Schmidt L, Mackintosh A, et al. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ: British Medical Journal 2002;**324**(7351):1417.

7 Ader DN. Developing the patient-reported outcomes measurement information system (PROMIS). Medical Care 2007;**45**(5):S1-2.

8 Dawson J, Fitzpatrick R, Murray D, et al. Questionnaire on the perceptions of patients about total knee replacement, Journal of Bone & Joint Surgery, British Volume 1998;**80**(1):63-9.
9 Dawson J, Doll H, Fitzpatrick R, et al. The routine use of patient reported outcome measures in healthcare settings. BMJ: British Medical Journal (Clinical Research ed.) 2009;**340**:c186.
10 McGrail K, Bryan S, Davis J. Let's all go to the PROM: the case for routine patient-reported outcome measurement in Canadian healthcare. HealthcarePapers 2011;**11**(4):8-18.
11 Devlin NJ, Appleby J, Buxton M. *Getting the most out of PROMs: putting health outcomes at the heart of NHS decision-making*. King's Fund, London, 2010.
12 Basch E. Standards for patient-reported outcome-based performance measures standards for patient-reported outcome-based performance measures viewpoint. JAMA: The Journal of Medical Association 2013;**310**(2):139-40 doi: 10.1001/jama.2013.6855 [published Online First: Epub Date]|.
13 Makoul G, Clayman ML. An integrative model of shared decision making in medical encounters. Patient Education and Counseling 2006;**60**(3):301-12.
14 Elwyn G, Edwards A, Kinnersley P. Shared decision-making in primary care: the neglected second half of the consultation. The British Journal of General Practice 1999;**49**(443):477-82.
15 Elwyn G, Edwards A, Kinnersley P, et al. Shared decision making and the concept of equipoise: the competences of involving patients in healthcare choices. The British Journal of General Practice 2000;**50**(460):892-9.
16 Edwards A, Elwyn G, Hood K, et al. Patient-based outcome results from a cluster randomized trial of shared decision making skill development and use of risk communication aids in general practice. Family Practice 2004;**21**(4):347-54.
17 Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. BMJ: British Medical Journal 2002;**324**(7341):827.
18 Stiggelbout A, Weijden T, Wit MD, et al. Shared decision making: really putting patients at the centre of healthcare. BMJ: British Medical Journal 2012;**344**:e256.
19 Elwyn G, Hutchings H, Edwards A, et al. The OPTION scale: measuring the extent that clinicians involve patients in decision-making tasks. Health Expectations 2005;**8**(1):34-42.
20 Coulter A, Collins A. *Making shared decision-making a reality. No decision about me without me*. The King's Fund, London, 2011.
21 Gravel K, Légaré F, Graham ID. Barriers and facilitators to implementing shared decision-making in clinical practice: a systematic review of health professionals' perceptions. Implementation Science 2006;**1**(1):16.
22 Elwyn G, Rix A, Holt T, et al. Why do clinicians not refer patients to online decision support tools? Interviews with front line clinics in the NHS. BMJ Open 2012;**2**(6) doi: 10.1136/bmjopen-2012-001530 [published Online First: Epub Date]|.
23. Elwyn G, Lloyd A, Joseph-Williams N, et al. Option Grids: shared decision making made easier. Patient Education and Counseling 2013;**90**:207-12.
24 Elwyn G, Lloyd A, Williams NJ, et al. Shared decision-making in a multidisciplinary head and neck cancer team: a case study of developing Option Grids. International Journal of Person Centered Medicine 2012;**2**(3):421-6.
25 Thomson R, Kinnersley P, Barry M. Shared decision making: a model for clinical practice. Journal of General Internal Medicine 2012;**27**(10):1361-7.

26 March L, Irwig L, Schwarz J, et al. n of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis. BMJ: British Medical Journal 1994;**309**(6961):1041-6.
27 Lillie EO, Patay B, Diamant J, et al. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Personalized Medicine 2011;**8**(2):161-73.
28. Moore A, Derry S, Eccleston C, et al. Expect analgesic failure; pursue analgesic success. BMJ: British Medical Journal 2013;**346**:f2690.

Capítulo 17 Críticas a la medicina basada en la evidencia

¿Qué tiene de malo la MBE cuando se usa mal?

Este nuevo capítulo es necesario porque la medicina basada en la evidencia (MBE) hace mucho que pasó años dorados. Hay, con gran acierto, un conjunto creciente de autores que plantean críticas legítimas a los supuestos y planteamientos centrales de la MBE. También hay una serie un poco mayor de críticas mal documentadas, así como una zona gris de literatura «anti-MBE» que contiene cierto grado de verdad, pero que es parcial y está mal argumentada. En este capítulo se pretende exponer las críticas legítimas y presentar al lector interesado argumentos más profundos.

Para documentar este capítulo, me he basado en varias fuentes: un breve artículo ampliamente citado de Spence[1], que es columnista del BMJ y un médico general con un gran sentido común, un libro de Timmermans y Berg[2] titulado *The Gold Standard: The challenge of evidence-based medicine and standardization in health care,* un artículo de Timmermans y Mauck[3] sobre las promesas y peligros de la MBE, una reflexión «pasados 20 años» de algunos gurús de la MBE[4], el libro de Goldacre[5] *Bad Pharma* y algunos materiales adicionales sobre la elaboración de políticas basadas en la evidencia que se citan en la sección «¿Por qué la elaboración de políticas basadas en la evidencia es tan difícil de lograr?».

Lo primero que debemos tener claro es la distinción entre la MBE cuando se usa mal (esta sección) y la MBE cuando se usa bien (sección siguiente). Como aperitivo de esta sección, se reproducen dos párrafos del prólogo de este libro, escrito para la primera edición en el año 1995 y que aún no se ha modificado en esta quinta edición:

> *Muchas de las descripciones dadas por los cínicos de lo que es la medicina basada en la evidencia (la glorificación de las cosas que pueden medirse sin tener en cuenta la utilidad o exactitud de lo que se mide, la aceptación acrítica de los datos numéricos publicados, la preparación de guías que abarcan todo por autoproclamados «expertos» que están alejados de la medicina real, la degradación de la libertad clínica por la imposición de protocolos clínicos rígidos y dogmáticos, y el exceso de confianza en análisis económicos simplistas, inapropiados, y a menudo incorrectos) en realidad*

son críticas contra aquello que el movimiento de la medicina basada en la evidencia combate y no de lo que representa.

Sin embargo, no quiero ser considerada una evangelista de la medicina basada en la evidencia. Creo que la ciencia de la búsqueda, evaluación y aplicación de los resultados de la investigación médica puede (y a menudo lo hace) lograr que la asistencia del paciente sea más objetiva, más lógica y más coste-efectiva. Si yo no creyese en esto, no dedicaría tanto tiempo enseñándola e intentando practicarla como médico general. Sin embargo, creo que cuando se aplica en vacío (es decir, sin sentido común y sin tener en cuenta las circunstancias y prioridades individuales de la persona a la que se ofrece tratamiento o la compleja naturaleza de la práctica clínica y la elaboración de políticas), la toma de decisiones «basada en la evidencia» es un proceso reduccionista que puede ser verdaderamente perjudicial.

A continuación, se expondrán estas cuestiones con más detalle. ¿Cómo es la «MBE mal utilizada»?

En primer lugar, la MBE de mala calidad emplea cifras derivadas de estudios poblacionales, pero no plantea preguntas previas sobre la procedencia de esas cifras (o estudios). Quien haya pasado tiempo en una planta de hospital o en una consulta de medicina general sabrá cuál es el tipo de persona que tiende a hacer esto: un médico locuaz, adepto de la técnica que parece conocer la literatura y cómo acceder a ella (tal vez a través de aplicaciones de su tableta de último modelo) y que siempre parece tener un NNT (número necesario a tratar) o una odds ratio en la punta de los dedos. Sin embargo, a esta persona verborreica se le da peor justificar por qué debería preferirse *este* conjunto de cifras «basadas en la evidencia» en lugar de otro. Su evidencia, por ejemplo, puede proceder de un único ensayo en lugar de un metaanálisis reciente de alta calidad de todos los ensayos disponibles. Los «expertos» locuaces y autoproclamados en MBE tienden a ser irreflexivos (es decir, no dedican mucho tiempo a pensar profundamente sobre las cosas) y pocas veces tienen una opinión *crítica* de los números que citan. Es posible que no compartan los argumentos sobre los criterios de valoración indirectos descritos en la página 81.

La MBE de mala calidad considera que el mundo de la evidencia publicada es igual al mundo de las necesidades del paciente. Por lo tanto, incurre en dos falacias: asume que si, por ejemplo, existe un ensayo controlado aleatorizado (ECA) que haya evaluado un tratamiento para una «enfermedad», dicha enfermedad es necesariamente un problema médico real que requiere tratamiento, y también asume que si no existe evidencia «metodológicamente sólida» sobre un tema, ese tema no es importante. Esto da lugar a un sesgo significativo. La base de evidencia se incrementaría en las enfermedades donde existan perspectivas de ganancias para la industria farmacéutica y de dispositivos médicos, como la detección, monitorización y control de los factores de riesgo de enfermedades cardiovasculares[6], el desarrollo y evaluación de nuevas clases farmacológicas para

la diabetes[7] o la creación y el tratamiento de cuadros que no son verdaderas enfermedades, como el «deseo sexual femenino hipoactivo»[8]). La evidencia también se acumulará en las enfermedades que el gobierno opta por reconocer y priorizar para la investigación financiada con fondos públicos, pero no se incrementará (o lo hará mucho más despacio) en las enfermedades «Cenicienta» que la industria y/o el gobierno consideran poco importantes, difíciles de clasificar o no «médicas», como la multimorbilidad[9], la actividad física en la prevención cardiovascular[10], la violencia doméstica[11] o la fragilidad relacionada con la edad[12].

La MBE de mala calidad prácticamente no tiene en cuenta la perspectiva del paciente ni tiene en consideración la importancia del criterio clínico. Como he señalado en la sección «La perspectiva del paciente», el «mejor» tratamiento no es necesariamente el que se ha demostrado que es el más eficaz en ECA, sino el que se ajusta a un conjunto particular de circunstancias concretas y coincide con las preferencias y prioridades del paciente.

Por último, la MBE de mala calidad se basa en las investigaciones defectuosas, por ejemplo, aquellas en las cuales se han utilizado estrategias de muestreo inadecuadas, tamaños muestrales injustificados, comparadores inapropiados, trucos estadísticos, etcétera. En el capítulo 6 se indican algunas formas específicas en las cuales la investigación (y la forma en que se presenta) puede inducir a error. Aunque las personas que actúan de esta manera a menudo se proclaman miembros de la comunidad de la MBE (p. ej., sus trabajos pueden incluir «basado en la evidencia» en el título), los miembros más académicos de esa comunidad contradirían enérgicamente estas alegaciones.

¿Qué tiene de malo la MBE cuando se usa bien?

Aunque la parte médica que hay en mí se preocupa por el uso inadecuado de la MBE, mi vertiente académica se interesa más por sus limitaciones cuando se emplea bien. Esto se debe a que hay buenas razones filosóficas por las cuales la MBE nunca será la fuente de todo el conocimiento.

Una crítica importante de la MBE, señalada por Timmermans y Berg en su libro, es el grado en que la MBE es un método formalizado para imponer un grado injustificable de estandarización y control de la práctica clínica. Estos autores argumentan que en el mundo clínico moderno, la MBE puede equipararse más o menos con la producción e implementación de guías de práctica clínica. «Sin embargo», alegan (pág. 3), «pocas veces se dispone de tal evidencia para abarcar todos los momentos de toma de decisiones de una guía. Para llenar las lagunas e interpretar las afirmaciones contradictorias que pudieran existir en la literatura, se necesitan pasos menos objetivos adicionales [como métodos de consenso] para crear una guía»[2].

Debido a estas lagunas (a veces sutiles) en la base de investigación, Timmermans y Berg sostienen que una guía «basada en la evidencia» no suele estar tan basada en la evidencia como parece. Sin embargo, la *formalización* de la evidencia en guías, que luego pueden osificarse en protocolos o programas informatizados de apoyo

a la toma de decisiones, proporciona un nivel injustificado de significación (y a veces de coerción) a la guía. Las asperezas se liman, las oquedades se rellenan y las recomendaciones resultantes empiezan a adquirir importancia bíblica.

Un efecto secundario desagradable de esta osificación es que la mejor evidencia *de ayer* lastra las guías y las vías clínicas *de hoy*. Un ejemplo es la disminución de la glucemia en la diabetes tipo 2. Durante muchos años, la hipótesis «basada en la evidencia» era que, cuanto más intensivo fuese el control de la glucemia de una persona, mejores serían los resultados. Sin embargo, más recientemente un gran metaanálisis demostró que el control intensivo de la glucosa no proporcionaba ningún beneficio respecto al control moderado y además se asociaba con un aumento del doble de la incidencia de hipoglucemia grave[13]. Sin embargo, la actuación de los médicos generales de Reino Unido todavía se regía mediante un esquema denominado *Quality and Outcomes Framework* (QOF) para intentar un control intensivo de la glucemia *después* de que la publicación de ese metaanálisis demostrase una relación beneficio-perjuicio adversa[14]. Esto se debe a que se necesita tiempo para que la práctica y la política se pongan al día con la evidencia, pero la existencia del QOF, introducido para que la asistencia estuviese más basada en la evidencia, en realidad tuvo el efecto de hacerla *menos* basada en la evidencia.

Tal vez la mayor crítica contra la MBE es que, si se utiliza mal, desprecia la perspectiva del paciente sobre la enfermedad para centrarse en el efecto promedio en una muestra poblacional o en el conjunto de datos de años de vida ajustados por calidad (AVAC) (v. cap. 11) calculados por un estadístico médico. Algunos autores que escriben sobre MBE se muestran entusiasmados por el uso de un enfoque de algoritmos para incorporar la perspectiva del paciente a una opción terapéutica basada en la evidencia. En la práctica, esto suele resultar imposible ya que, como he señalado en la sección «La perspectiva del paciente», las experiencias de los pacientes son historias complejas que no se pueden reducir a un algoritmo de decisiones de tipo sí/no (o «con tratamiento, sin tratamiento»).

La imposición (real) de una asistencia estandarizada reduce la capacidad del médico de responder a las cuestiones idiosincrásicas e inmediatas que surgen en una consulta particular. La esencia más íntima del enfoque de la MBE es utilizar una media poblacional (o más exactamente, un promedio de una muestra representativa) para documentar la toma de decisiones para ese paciente. Sin embargo, como como muchos otros autores antes que yo han señalado, un paciente no es una media ni una mediana, sino un individuo, cuya enfermedad inevitablemente tiene características únicas e inclasificables. Una estandarización excesiva no sólo hace que la asistencia ofrecida coincida menos con las necesidades individuales, sino que también provoca una pérdida de las aptitudes del médico, por lo que pierde la capacidad de adaptar y personalizar la asistencia (o, en el caso de los médicos recién formados, no logra obtener dicha capacidad de entrada).

En palabras de Spence[1]: «la evidencia genera un sentido del absolutismo, pero el absolutismo se debe temer absolutamente. Nuestra medicina reduccionista basada en algoritmos ha dado lugar a la idea de que no se puede ir en contra de la evidencia, con la consiguiente polifarmacia irreflexiva, sobre todo en poblaciones

con enfermedades concurrentes. Como resultado, muchos miles de personas fallecen directamente por reacciones adversas a fármacos».

A continuación plantearé otro ejemplo. Recientemente realicé una investigación que me obligó a dedicar mucho tiempo a observar a médicos residentes en un servicio de urgencias. Descubrí que siempre que un niño acudía con una lesión, el médico residente cumplimentaba una serie de preguntas en la historia clínica electrónica del paciente. Estas preguntas se basaban en una guía basada en la evidencia para descartar lesiones no accidentales, pero debido a que los médicos residentes cumplimentaban las preguntas en todos los niños, me parecía que las sospechas que podrían haber tenido en el caso de cualquier niño *particular* estaban ausentes. Este enfoque estandarizado contrastaba con mis propios días de médico residente hace 30 años, cuando no teníamos guías, pero dedicábamos algo de tiempo a evaluar y perfeccionar nuestras sospechas.

Otra de las preocupaciones acerca de la «MBE bien utilizada» es la gran cantidad de orientación y consejos basados en la evidencia que existe en la actualidad. Como señalé en la sección «El gran debate sobre las guías», las guías necesarias para tratar la cantidad de pacientes atendidos en una guardia típica de 24 h abarcan más de 3.000 páginas y requerirían más de una semana para que un médico las leyese[15]. Y eso sin incluir las indicaciones de punto de asistencia que sugerían otras intervenciones basadas en la evidencia (p. ej., gestión de los factores de riesgo) en los pacientes atendidos fuera de un entorno de urgencias. Por ejemplo, siempre que veo a un paciente de 16-25 años en la consulta, un indicador emergente me recuerda que ofrezca el cribado de clamidias. Uno de mis artículos cualitativos publicado conjuntamente con Swinglehurst[16] ha demostrado que tales indicaciones son muy perjudiciales para la dinámica de la relación médico-paciente.

Una crítica más filosófica de la MBE es que se basa en una versión simplista e ingenua de lo que es el conocimiento. La suposición es que el conocimiento se puede equiparar con los «hechos» derivados de los estudios de investigación que pueden formalizarse en guías y traducirse (es decir, implementarse por profesionales y elaboradores de políticas). Sin embargo, como ya he dicho en otro lugar, el conocimiento es una fiera compleja e imprevisible[17]. Por un lado, sólo parte del conocimiento se puede considerar algo que una persona puede asimilar como un «hecho»; existe otro nivel de conocimiento que es *colectivo,* es decir, socialmente compartido e integrado a nivel organizacional[18]. En palabras de Tsoukas y Vladimirou[19]:

> *El conocimiento es una mezcla fluida de experiencias enmarcadas, valores, información contextual y una visión experta que proporciona un marco para evaluar e incorporar nuevas experiencias e información. Se origina y se aplica en la mente de los conocedores. En las organizaciones, a menudo se integra no sólo en los documentos o repositorios, sino también en las rutinas, procesos, prácticas y normas organizacionales.*

Gabbay y May[20] ilustraron este elemento colectivo de conocimiento en su estudio que mencioné brevemente en la sección «¿Cómo se puede ayudar a garantizar

que se siguen las guías basadas en la evidencia?» de la página 138. Aunque estos investigadores, que observaron a médicos generales en acción durante varios meses, nunca vieron a los médicos consultar las guías directamente, sí los observaron comentar y negociar estas guías entre sí y también actuar de un modo que demostró que habían absorbido e incorporado de alguna manera «por ósmosis» los componentes clave de muchas guías basadas en la evidencia. Estos elementos incorporados colectivamente y socialmente compartidos de guías son lo que Gabbay y May denominan guías mentales *(mindlines)*.

Los «hechos» que poseen los individuos (p. ej., el hallazgo de una investigación que una persona ha descubierto en una búsqueda exhaustiva de la literatura) pueden colectivizarse mediante diversos mecanismos, incluidos los esfuerzos dirigidos a hacerlos relevantes para sus colegas (oportunos, destacados, factibles), legítimos (creíbles, fiables, razonables) y accesibles (disponibles, comprensibles, asimilables) y para tener en cuenta los puntos de partida (suposiciones, visiones del mundo, prioridades) de un público en particular.

Estos mecanismos son elementos de la ciencia de la traducción del conocimiento (un tema importante que está más allá del alcance de este libro)[17,20-22]. El punto clave aquí es que presentar la MBE puramente como la secuencia de tareas individuales establecidas en los capítulos anteriores de este libro es una descripción demasiado simplista. Recomiendo a los lectores que se sientan familiarizados con los fundamentos de la MBE que consulten la literatura sobre estas dimensiones más amplias del conocimiento.

¿Por qué la elaboración de políticas basadas en la evidencia es tan difícil de lograr?

La principal crítica que algunas personas realizan a la MBE es que no logra trasladar la evidencia de un modo simple y lógico a las políticas. La razón de por qué las políticas no fluyen de manera sencilla y lógica desde la evidencia de la investigación es que hay muchos otros factores implicados.

Pongamos como ejemplo la cuestión de los tratamientos para la infertilidad financiados con fondos públicos. Podemos recopilar una cantidad de pruebas tan alta como una casa para demostrar que la intervención X proporciona una tasa de bebés sanos del Y% en mujeres con características (como la edad o las enfermedades concurrentes) Z, pero esto no rebajará el calor del debate existente sobre la decisión de autorizar el tratamiento de la infertilidad en el marco de un presupuesto de asistencia sanitaria limitado. Ésta fue la cuestión abordada por el foro de elaboración de políticas *Primary Care Trust*, al cual asistí recientemente, que tuvo que sopesar esta decisión frente a opciones competidoras (apoyo de extensión para un primer episodio de psicosis y enfermera comunitaria especializada en diabetes para epilepsia). Los miembros del foro no ignoraban la evidencia (había tanta evidencia en los artículos informativos que me enviaron que el mensajero no logró introducirlos en mi buzón), pero la decisión final se basó en los valores en lugar de basarse en la evidencia. Y como muchos han señalado, la elaboración de políticas tiene tanto

que ver con el esfuerzo por resolver los conflictos de valores en contextos locales o nacionales particulares como con aplicar la evidencia en la práctica[23].

Dicho de otro modo, el proceso de la elaboración de políticas no se puede considerar como una versión «macro» de la secuencia representada en la sección 1.1 («convertir nuestra necesidades de información en preguntas con respuesta…», etc.). Al igual que otros procesos que se engloban en las «políticas» (con «p» minúscula), la elaboración de políticas consiste fundamentalmente en persuadir a alguno de los altos cargos que toman decisiones de la superioridad de una línea de actuación respecto a otra. Este modelo del proceso de elaboración de políticas tiene un respaldo sólido de estudios de investigación, lo que sugiere que en su núcleo existen componentes de imprevisibilidad, ambigüedad y la posibilidad de interpretaciones alternativas de la «evidencia»[23,24].

Es posible que la búsqueda para lograr que la elaboración de políticas esté «totalmente basada en la evidencia» en realidad no sea un objetivo deseable, ya que este criterio podría devaluar el debate democrático sobre las cuestiones éticas y morales que plantean las opciones políticas. El manifiesto del partido laborista británico de 2005 afirmaba que «lo que importa es lo que funciona». Sin embargo, lo que importa, sin duda, no es sólo lo que «funciona», sino lo que es apropiado en cada circunstancia y lo que la sociedad acuerda que es el objetivo deseable global. Deborah Stone, en su libro *Policy Paradox,* argumenta que gran parte del proceso de elaboración de políticas implica debates sobre los valores que se hacen pasar por debates sobre hechos y datos. En sus propias palabras: «la esencia de la elaboración de políticas en las comunidades políticas [es] la lucha por las ideas. Las ideas están en el centro de todo conflicto político… Cada idea es un argumento o, más exactamente, una colección de argumentos a favor de diferentes formas de ver el mundo»[25].

Dobrow y cols.[26] han escrito uno de los artículos teóricos más útiles sobre el uso de la evidencia en la elaboración de políticas de asistencia sanitaria. Estos autores distinguen la orientación filosófico-normativa (que existe una realidad objetiva que debe descubrirse y que un fragmento de «evidencia» puede considerarse «válido» y «fiable» con independencia del contexto en el que se va a utilizar) de la orientación práctico-operativa, en la cual la evidencia se define en relación con un contexto de toma de decisiones específicas, nunca es estática y se caracteriza por la emergencia, la ambigüedad y su carácter incompleto. Desde un punto de vista práctico-operativo, la evidencia de la investigación se basa en diseños (como los ensayos aleatorizados) que eliminan explícitamente del estudio los «contaminantes» contextuales y que, por lo tanto, ignoran los múltiples determinantes, complejos e interactivos, de la salud. De ello se desprende que una intervención compleja que «funciona» en un entorno y en un momento dado no funcionará necesariamente en un entorno diferente en un momento distinto y que una que demuestra ser «coste-efectiva» en un entorno no será necesariamente rentable en un entorno diferente. Muchos de los argumentos planteados sobre la MBE en los últimos años han abordado precisamente esta controversia sobre la naturaleza del conocimiento.

El cuestionamiento de la naturaleza de la evidencia y, de hecho, el cuestionamiento del propio conocimiento basado en la evidencia son una forma de terminar

un libro de texto introductorio básico sobre MBE que da un poco de miedo porque la mayoría de los capítulos previos de este libro asumen lo que Dobrow llamaría una orientación filosófico-normativa. Mi consejo es el siguiente: si el lector es un estudiante que trata de aprobar sus exámenes o un médico que intenta mejorar su trabajo a la cabecera de los pacientes, y si se siente desconcertado por las incertidumbres que he planteado en esta última sección, puede ignorarlas sin problema hasta que esté activamente involucrado en la elaboración de políticas. Sin embargo, si su carrera está en la etapa en la cual forma parte de los órganos de toma de decisiones y trata de averiguar la respuesta a la pregunta planteada en el título de esta sección, le sugeriría que consultase algunos de los artículos y libros que aparecen en la bibliografía de esta sección. Esté atento a la próxima generación de investigación sobre MBE, que se ocupa de forma creciente de los aspectos más difusos y discutibles de este importante tema.

Bibliografía

1 Spence D. Why evidence is bad for your health. BMJ: British Medical Journal 2010;**341**:c6368.
2 Timmermans S, Berg M. *The gold standard: the challenge of evidence-based medicine and standardization in health care*. Philadelphia: Temple University Press, 2003.
3 Timmermans S, Mauck A. The promises and pitfalls of evidence-based medicine. Health Affairs 2005;**24**(1):18-28.
4 Agoritsas T, Guyatt GH. Evidence-based medicine 20 years on: a view from the inside. The Canadian Journal of Neurological Sciences 2013;**40**(4):448-9.
5 Goldacre B. *Bad pharma: how drug companies mislead doctors and harm patients*. London, Fourth Estate, Random House Digital Inc, 2013.
6 Saukko PM, Farrimond H, Evans PH, et al. Beyond beliefs: risk assessment technologies shaping patients' experiences of heart disease prevention. Sociology of Health & Illness 2012;**34**(4):560-75.
7 Davis C, Abraham J. The socio-political roots of pharmaceutical uncertainty in the evaluation of 'innovative'diabetes drugs in the European Union and the US. Social Science & Medicine 2011;**72**(9):1574-81.
8 Jutel A. Framing disease: the example of female hypoactive sexual desire disorder. Social Science & Medicine 2010;**70**(7):1084-90.
9 Lugtenberg M, Burgers JS, Clancy C, et al. Current guidelines have limited applicability to patients with comorbid conditions: a systematic analysis of evidence-based guidelines. PloS One 2011;**6**(10):e25987.
10 Bull FC, Bauman AE. Physical inactivity: the "Cinderella" risk factor for noncommunicable disease prevention. Journal of Health Communication 2011;**16**(Suppl. 2):13-26.
11 Garcia-Moreno C, Watts C. Violence against women: an urgent public health priority. Bulletin of the World Health Organization 2011;**89**(1):2.
12 Clegg A, Young J, Iliffe S, et al. Frailty in elderly people. The Lancet 2013;**381**:752-62.
13 Boussageon R, Bejan-Angoulvant T, Saadatian-Elahi M, et al. Effect of intensive glucose lowering treatment on all cause mortality, cardiovascular death, and microvascular events in type 2 diabetes: meta-analysis of randomised controlled trials. BMJ: British Medical Journal 2011;**343**:d4169.

14 Calvert M, Shankar A, McManus RJ, et al. Effect of the quality and outcomes framework on diabetes care in the United Kingdom: retrospective cohort study. BMJ: British Medical Journal 2009;**338**:b1870.
15 Allen D, Harkins K. Too much guidance? The Lancet 2005;**365**(9473):1768.
16 Swinglehurst D, Roberts C, Greenhalgh T. Opening up the 'black box' of the electronic patient record: a linguistic ethnographic study in general practice. Communication & Medicine 2011;**8**(1):3-15.
17 Greenhalgh T. What is this knowledge that we seek to "exchange"? The Milbank Quarterly 2010;**88**(4):492-9 doi: 10.1111/j.1468-0009.2010.00610.x[published Online First: Epub Date].
18 Contandriopoulos D, Lemire M, DENIS JL, et al. Knowledge exchange processes in organizations and policy arenas: a narrative systematic review of the literature. Milbank Quarterly 2010;**88**(4):444-83.
19 Tsoukas H, Vladimirou E. What is organizational knowledge? Journal of Management Studies 2001;**38**(7):973-83.
20 Gabbay J, May Al. Evidence based guidelines or collectively constructed "mindlines?" Ethnographic study of knowledge management in primary care. BMJ: British Medical Journal 2004;**329**(7473):1013.
21 Greenhalgh T, Wieringa S. Is it time to drop the 'knowledge translation'metaphor? A critical literature review. Journal of the Royal Society of Medicine 2011;**104**(12):501-9 doi: 10.1258/jrsm.2011.110285[published Online First: Epub Date].
22 Graham ID, Logan J, Harrison MB, et al. Lost in knowledge translation: time for a map? Journal of Continuing Education in the Health Professions 2006;**26**(1):13-24.
23 Greenhalgh T, Russell J. Evidence-based policymaking: a critique. Perspectives in Biology and Medicine 2009;**52**(2):304-18.
24 Scheel I, Hagen K, Oxman A. The unbearable lightness of healthcare policy making: a description of a process aimed at giving it some weight. Journal of Epidemiology and Community Health 2003;**57**(7):483-7.
25 Stone DA. *Policy paradox: the art of political decision making*. New York: WW Norton, 1997.
26 Dobrow MJ, Goel V, Upshur R. Evidence-based health policy: context and utilisation. Social Science & Medicine 2004;**58**(1):207-17.

Apéndice 1 Listas de comprobación para encontrar, evaluar e implementar la evidencia

A menos que se indique lo contrario, estas listas de comprobación se pueden aplicar a los ensayos controlados aleatorizados, otros ensayos clínicos controlados, estudios de cohortes, estudios de casos y controles, o cualquier otra evidencia proveniente de la investigación.

¿Está mi práctica basada en la evidencia? Lista de comprobación sensible al contexto para encuentros clínicos individuales (v. cap. 1)

1. ¿He identificado y priorizado los problemas clínicos, psicológicos, sociales y de otro tipo, teniendo en cuenta la perspectiva del paciente?
2. ¿He realizado una exploración suficientemente competente y completa para establecer la probabilidad de diagnósticos diferenciales?
3. ¿He considerado los problemas y factores de riesgo adicionales que pueden necesitar atención oportuna?
4. En caso necesario, ¿he buscado evidencia (en revisiones sistemáticas, guías, ensayos clínicos y otras fuentes) relativa a los problemas?
5. ¿He evaluado y he tenido en cuenta la integridad, la calidad y la fuerza de la evidencia?
6. ¿He aplicado una evidencia válida y relevante para este conjunto particular de problemas de manera que sea a la vez científicamente justificada e intuitivamente razonable?
7. ¿He presentado los pros y los contras de las diferentes opciones para el paciente de manera que él lo pueda entender y he incorporado las preferencias del paciente en la recomendación final?
8. ¿He dispuesto una revisión, recordatorio, derivación u otra asistencia adicional en caso de ser necesario?

Lista de comprobación para las búsquedas (v. cap. 2)

1. Decida el propósito de la búsqueda: consulta de la literatura, buscar una respuesta a una pregunta clínica o una revisión exhaustiva (p. ej., antes de realizar un trabajo de investigación), y diseñe su estrategia de búsqueda consecuente (sección «¿Qué estamos buscando?»).

2. Busque el nivel de evidencia más alto posible (sección «Jerarquías de los niveles de evidencia»). Por ejemplo, las fuentes sintetizadas de alta calidad (p. ej., revisiones sistemáticas y resúmenes y síntesis basados en la evidencia, como Clinical Evidence o guías NICE, sección «Fuentes sintetizadas: sistemas, resúmenes y síntesis») representan un nivel de evidencia muy alto.
3. Para mantenerse al tanto de los nuevos avances, utilice sinopsis como POEMS (*patient-oriented evidence that matters*), *ACP Journal Club* o revista *Evidence Based Medicine* (sección «Fuentes preevaluadas: sinopsis de revisiones sistemáticas y estudios primarios»).
4. Familiarícese con los recursos especializados de su propio campo y utilícelos habitualmente (sección «Recursos especializados»).
5. Al buscar investigaciones primarias en la base de datos Medline, se aumentará en gran medida la eficiencia de la búsqueda si se realizan dos búsquedas generales y luego se combinan, o si se utilizan herramientas, como la función *set limit* (limitar) o *clinical queries* (consultas clínicas) (sección «Estudios primarios: desentrañando la selva»).
6. Una manera muy potente de identificar publicaciones recientes sobre un tema es el encadenamiento de citas *(citation chaining)* de un artículo más antiguo (es decir, utilizar una base de datos electrónica especial para encontrar qué artículos posteriores han citado el más antiguo (sección «Estudios primarios: desentrañando la selva»).
7. Los motores de búsqueda federada, como TRIP o SUMsearch, buscan en múltiples recursos simultáneamente y son gratis (sección «Sistema de ventanilla única: motores de búsqueda federada»).
8. Las fuentes humanas (bibliotecarios expertos, expertos en el campo) son un componente importante de una búsqueda exhaustiva (sección «Fuentes de ayuda y preguntas a conocidos»).
9. Para mejorar su habilidad y confianza en la búsqueda, pruebe un curso de autoestudio en línea (sección «Tutoriales en línea para una búsqueda eficaz»).

Lista de comprobación para determinar de qué trata un artículo (v. cap. 3)

1. ¿Por qué se realizó el estudio (qué pregunta clínica abordó)?
2. ¿Qué tipo de estudio se llevó a cabo?
   - ¿Investigación primaria (experimento, ensayo controlado aleatorizado, otro ensayo clínico controlado, estudio de cohortes, estudio de casos y controles, estudio transversal, estudio longitudinal, caso aislado o serie de casos)?
   - ¿Investigación secundaria (revisión simple, revisión sistemática, metaanálisis, análisis de decisiones, desarrollo de guías o análisis económico)?
3. ¿Fue el diseño del estudio apropiado para el campo general de investigación abordado (tratamiento, diagnóstico, cribado, pronóstico y causalidad)?
4. ¿El estudio cumplió los estándares previstos de ética y control?

Lista de comprobación para la sección de métodos de un artículo (v. cap. 4)

1. ¿Era el estudio original?
2. ¿Qué pacientes incluye el estudio?
   - ¿Cómo se reclutó a los participantes?
   - ¿Quién fue incluido y quién fue excluido del estudio?
   - ¿Se estudió a los participantes en circunstancias de la «vida real»?
3. ¿El diseño del estudio era acertado?
   - ¿Qué intervención u otra actuación se estaba evaluando?
   - ¿Qué resultado(s) se midió (midieron), y cómo?
4. ¿El estudio se controló de manera adecuada?
   - Si era un ensayo aleatorizado, ¿fue la asignación verdaderamente aleatoria?
   - Si era un estudio de cohortes, de casos y controles, u otro estudio comparativo no aleatorizado, ¿eran los controles apropiados?
   - ¿Fueron los grupos comparables en todos los aspectos importantes a excepción de la variable que se estaba estudiando?
   - ¿Fue la evaluación del resultado (o, en un estudio de casos y controles, la definición de caso) «ciega»?
5. ¿Fue el estudio lo suficientemente grande y se continuó durante el tiempo suficiente, y fue el seguimiento lo bastante completo para que los resultados fuesen creíbles?

Lista de comprobación para los aspectos estadísticos de un artículo (v. cap. 5)

1. ¿Han planteado los autores correctamente el escenario?
   - ¿Han determinado si sus grupos son comparables y, si es necesario, han realizado los ajustes en función de las diferencias iniciales?
   - ¿Qué tipo de datos se han recogido? ¿Se han utilizado las pruebas estadísticas apropiadas?
   - Si las pruebas estadísticas usadas en el artículo son poco claras, ¿por qué las escogieron los autores?
   - ¿Se han analizado los datos de acuerdo con el protocolo original del estudio?
2. Datos pareados, colas y valores atípicos.
   - ¿Se realizaron las pruebas pareadas con datos pareados?
   - ¿Se ha realizado una prueba de dos colas siempre que el efecto de una intervención pudiese haber sido negativo?
   - ¿Se analizaron los valores atípicos con sentido común y se realizaron los ajustes estadísticos apropiados?
3. Correlación, regresión y causalidad.
   - ¿Se ha distinguido la correlación de la regresión, y se ha calculado e interpretado adecuadamente el coeficiente de correlación («valor *r*»)?
   - ¿Se han hecho suposiciones sobre la naturaleza y la dirección de la causalidad?

4. Probabilidad y confianza.
   - ¿Se han calculado e interpretado adecuadamente los «valores *p*»?
   - ¿Se han calculado los intervalos de confianza y se reflejan en las conclusiones de los autores?
5. ¿Han expresado los autores los efectos de una intervención en términos del beneficio o perjuicio probable que puede esperar un paciente individual, como:
   - reducción del riesgo relativo;
   - reducción del riesgo absoluto;
   - número necesario a tratar?

Lista de comprobación para el material proporcionado por un representante de una compañía farmacéutica (v. cap. 6)

Véase especialmente la tabla 6.1 para las preguntas sobre los ensayos aleatorizados basados en la declaración CONSORT.

1. ¿Se refiere este material a un tema que es de importancia clínica en mi práctica?
2. ¿Se ha publicado este material en revistas independientes revisadas por pares? ¿Se ha omitido cualquier evidencia significativa de esta presentación o se ha evitado su publicación?
3. ¿El material incluye evidencia de alto nivel como revisiones sistemáticas, metaanálisis o ensayos doble ciego controlados y aleatorizados frente al principal competidor del fármaco administrado en dosis óptima?
4. ¿Los ensayos o revisiones han abordado una pregunta clínica claramente centrada, importante y que pueda responderse, que refleje un problema relevante para los pacientes? ¿Proporcionan evidencia sobre seguridad, tolerancia, eficacia y precio?
5. ¿Todos los ensayos o metaanálisis definen la enfermedad que se va a tratar, los pacientes que se van a incluir, las intervenciones que se van a comparar y los resultados que deben evaluarse?
6. ¿El material proporciona evidencia directa de que el fármaco ayudará a mis pacientes a tener una vida más larga, más sana, más productiva y sin síntomas?
7. Si se ha utilizado una medida de resultado indirecta, ¿cuál es la evidencia de que es fiable, reproducible, sensible, específica, un verdadero factor predictivo de la enfermedad y que refleje rápidamente la respuesta al tratamiento?
8. ¿Los resultados del ensayo indican si (y cómo) la eficacia de los tratamientos difirió y si existía una diferencia en el tipo o la frecuencia de reacciones adversas? ¿Los resultados se expresan en términos de números necesarios a tratar, y son significativos desde los puntos de vista clínico y estadístico?
9. Si el representante ha aportado grandes cantidades de material, ¿qué tres artículos proporcionan la evidencia más fuerte de las afirmaciones de la empresa?

Lista de comprobación para un artículo que describe un estudio de una intervención compleja (v. cap. 7)

1. ¿Cuál es el problema para el que esta intervención compleja se considera una posible solución?
2. ¿Qué se hizo en la fase de desarrollo de la investigación para documentar el diseño de la intervención compleja?
3. ¿Cuáles eran los componentes centrales y no centrales de la intervención?
4. ¿Cuál era el mecanismo de acción teórico de la intervención?
5. ¿Qué medidas de resultado se utilizaron, y eran sensibles?
6. ¿Cuáles fueron los hallazgos?
7. ¿Qué evaluación del proceso se realizó y cuáles fueron sus principales conclusiones?
8. Si los resultados fueron negativos, ¿hasta dónde se puede explicar esto por el fracaso a la hora de aplicar la intervención y/o por una optimización inadecuada de la misma?
9. Si los hallazgos variaron entre los diferentes subgrupos, ¿en qué medida lo han explicado los autores perfeccionando su teoría del cambio?
10. ¿Qué otras investigaciones son necesarias según los autores, y están justificadas?

Lista de comprobación para un artículo que pretende validar una prueba diagnóstica o de cribado (v. cap. 8)

1. ¿Es esta prueba potencialmente relevante para mi práctica?
2. ¿Se ha comparado la prueba con un verdadero patrón oro?
3. ¿Este estudio de validación incluye un espectro adecuado de participantes?
4. ¿Se ha evitado el sesgo de confirmación (verificación)?
5. ¿Se ha evitado el sesgo de observador?
6. ¿Se demostró que la prueba es reproducible en un mismo observador y entre distintos observadores?
7. ¿Cuáles son los parámetros de la prueba derivados de este estudio de validación?
8. ¿Se indicaron los intervalos de confianza para la sensibilidad, especificidad y otros parámetros de la prueba?
9. ¿Se ha obtenido un «rango de la normalidad» razonable a partir de estos resultados?
10. ¿Se ha puesto esta prueba en el contexto de otras pruebas posibles en la secuencia de diagnóstico de la enfermedad?

Lista de comprobación para una revisión sistemática o metaanálisis (v. cap. 9)

1. ¿Abordó la revisión una pregunta clínica importante?
2. ¿Se ha realizado una búsqueda exhaustiva en la(s) base(s) de datos apropiada(s) y se han explorado otras fuentes potencialmente importantes?

3. ¿Se evaluó la calidad metodológica (especialmente los factores que pueden predisponer al sesgo) y se ponderaron los ensayos en consonancia?
4. ¿Qué grado de sensibilidad tienen los resultados respecto al modo en que se ha realizado la revisión?
5. ¿Se han interpretado los resultados numéricos con sentido común y con la debida atención a los aspectos más generales del problema?

Lista de comprobación para un conjunto de guías clínicas (v. cap. 10)

1. ¿La preparación y publicación de esta guía conllevan un conflicto de intereses significativo?
2. ¿Las guías están relacionadas con un tema apropiado e indican claramente el objetivo del tratamiento ideal en términos de resultados de salud y/o coste?
3. ¿Intervino un especialista en la metodología de la investigación secundaria (p. ej., metaanalista)?
4. ¿Se han analizado todos los datos relevantes y las conclusiones de las guías concuerdan con los datos?
5. ¿La guía tiene en cuenta las variaciones de la práctica médica y otras áreas controvertidas (p. ej., una atención óptima en respuesta a una falta de financiación real o subjetiva)?
6. ¿Son las guías válidas y fiables?
7. ¿Son clínicamente relevantes, exhaustivas y flexibles?
8. ¿Tienen en cuenta lo que es aceptable, asequible y posible en la práctica para los pacientes?
9. ¿Incluyen recomendaciones para su propia difusión, aplicación y revisión periódica?

Lista de comprobación para un análisis económico (v. cap. 11)

1. ¿El análisis se basa en un estudio que responde a una pregunta clínica claramente definida sobre un tema de importancia económica?
2. ¿Desde qué punto de vista se consideran los costes y beneficios?
3. ¿Se ha demostrado que las intervenciones que se comparan son clínicamente eficaces?
4. ¿Son las intervenciones razonables y viables en los contextos en los que es probable que se vayan a aplicar?
5. ¿Qué método de análisis se utilizó, y era apropiado?
   - Si las intervenciones produjeron resultados idénticos ⇒ análisis de minimización de costes.
   - Si el resultado importante es unidimensional ⇒ análisis de coste-efectividad.

- Si el resultado importante es multidimensional ⇒ análisis de coste-utilidad.
- Si la ecuación coste-beneficio para esta enfermedad debe compararse con la ecuación coste-beneficio para otra enfermedad ⇒ análisis de coste-beneficio.
- Si un análisis de coste-beneficio fuese apropiado por lo demás, pero los valores de preferencia otorgados a diferentes estados de salud están cuestionados o es probable que cambien ⇒ análisis de coste-consecuencias.

6. ¿Cómo se midieron los costes y beneficios?
7. ¿Se tuvieron en cuenta los beneficios incrementales en lugar de los absolutos?
8. ¿Se ha dado prioridad al estado de salud de «aquí y ahora» respecto al del futuro lejano?
9. ¿Se realizó un análisis de sensibilidad?
10. ¿Se han usado en exceso las escalas agregadas «resumidas»?

Lista de comprobación para un artículo de investigación cualitativa (v. cap. 12)

1. ¿El artículo describe un problema clínico importante abordado mediante una pregunta claramente formulada?
2. ¿Fue apropiado usar un enfoque cualitativo?
3. ¿Cómo se seleccionaron (a) el contexto y (b) los individuos?
4. ¿Cuál fue la perspectiva del investigador, y se ha tenido en cuenta?
5. ¿Qué métodos utilizó el investigador para la recogida de datos y se describen con suficiente detalle?
6. ¿Qué métodos utilizó el investigador para analizar los datos y qué medidas de control de calidad se aplicaron?
7. ¿Los resultados son creíbles y, si es así, son clínicamente importantes?
8. ¿Qué conclusiones se extrajeron, y están justificadas por los resultados?
9. ¿Los resultados del estudio son transferibles a otros contextos clínicos?

Lista de comprobación para un artículo que describe investigaciones basadas en cuestionarios (v. cap. 13)

1. ¿Qué querían averiguar los investigadores, y fue un cuestionario el diseño de investigación más adecuado?
2. Si se disponía de un cuestionario «prefabricado» (es decir, publicado previamente y validado), ¿lo han utilizado los investigadores (y si no, por qué no)?
3. ¿Qué afirmaciones han hecho los investigadores sobre la validez del cuestionario (su capacidad para medir lo que ellos quieren medir) y su fiabilidad (su capacidad de dar resultados homogéneos a lo largo del

tiempo y en un mismo/o en distintos investigadores)? ¿Se justifican estas afirmaciones?
4. ¿El cuestionario estaba debidamente estructurado y presentado, y se redactaron los ítems adecuadamente para la sensibilidad del tema y el grado de conocimientos sanitarios de los encuestados?
5. ¿Se incluyeron instrucciones y explicaciones adecuadas?
6. ¿Se realizó una prueba piloto adecuada del cuestionario y se modificó la versión definitiva a la luz de los resultados piloto?
7. ¿Se seleccionó adecuadamente la muestra de posibles participantes y era lo bastante grande y representativa?
8. ¿Cómo se aplicó el cuestionario (p. ej., por correo postal, correo electrónico, teléfono) y se administró (autocumplimentación, cumplimentación ayudada por un investigador), y fueron estos enfoques apropiados?
9. ¿Se tuvieron en cuenta las necesidades de determinados subgrupos en el diseño y la administración del cuestionario? Por ejemplo, ¿qué se hizo para captar la perspectiva de los encuestados analfabetos o de quienes hablan un idioma diferente al del investigador?
10. ¿Cuál fue la tasa de respuesta, y por qué? Si la tasa de respuesta fue baja (<70%), ¿los investigadores han demostrado que no existen diferencias sistemáticas entre los respondedores y los no respondedores?
11. ¿Qué tipo de análisis se llevó a cabo sobre los datos del cuestionario, y era apropiado? ¿Hay alguna evidencia de «dragado de datos», es decir, análisis no basados en hipótesis?
12. ¿Cuáles fueron los resultados? ¿Eran definitivos (estadísticamente significativos), y también se describieron los resultados negativos y los no significativos?
13. ¿Se han interpretado de forma adecuada los datos cualitativos (p. ej., las respuestas de texto libre) (p. ej., utilizando un marco teórico explícito)? ¿Se han utilizado sensatamente las citas para ilustrar los hallazgos más generales en lugar de para adornar?
14. ¿Qué significan los resultados y han establecido los investigadores una relación adecuada entre los datos y sus conclusiones?

Lista de comprobación para un artículo que describe un estudio de mejora de la calidad (v. cap. 14)

1. ¿Cuál fue el contexto?
2. ¿Cuál fue el objetivo del estudio?
3. ¿Cuál era el mecanismo por el cual los autores esperaban mejorar la calidad?
4. ¿La iniciativa de mejora de la calidad prevista estaba basada en la evidencia?
5. ¿Cómo midieron el éxito los autores, y lo hicieron de forma razonable?

6. ¿Cuánto detalle se dio sobre el proceso de cambio, y qué perspectivas se puede extraer de esto?
7. ¿Cuáles fueron los principales hallazgos?
8. ¿Cuál fue la explicación para el éxito, el fracaso o la suerte dispar de la iniciativa? ¿Fue razonable?
9. A la luz de los resultados, ¿cuáles creen los autores que son los próximos pasos en el ciclo de mejora de la calidad a nivel local?
10. Según los autores, ¿cuáles eran las lecciones generalizables para otros equipos, y era esto razonable?

Lista de comprobación para las organizaciones sanitarias que trabajan dirigidas a una cultura basada en la evidencia para la toma de decisiones clínicas y de compra (v. cap. 15)

1. *Liderazgo:* ¿con qué frecuencia se ha comentado la información sobre efectividad o la medicina basada en la evidencia en las reuniones del consejo directivo en los últimos 12 meses? ¿El consejo ha dedicado tiempo para aprender sobre avances de efectividad clínica y de coste-efectividad?
2. *Inversión:* ¿qué recursos invierte la organización en la búsqueda y utilización de la información sobre efectividad clínica? ¿Hay un enfoque planificado para la promoción de la medicina basada en la evidencia que cuenta con los recursos y el personal adecuados?
3. *Políticas y guías:* ¿quién es responsable de recibir, actuar y monitorizar la implementación de políticas de orientación y recomendaciones políticas basadas en la evidencia, como las guías NICE o los Effective Health Care Bulletins? ¿Qué medidas se han adoptado basándose en cada una de estas publicaciones editadas hasta la fecha? ¿Las disposiciones garantizan que tanto los directivos como los médicos desempeñan su papel en el desarrollo y puesta en práctica de las guías?
4. *Formación:* ¿se ha proporcionado algún tipo de formación al personal de la organización (tanto clínico como no clínico) sobre la valoración y utilización de la evidencia de la efectividad para influir en la práctica clínica?
5. *Contratos:* ¿con qué frecuencia la información clínica y de coste-efectividad constituye una parte importante de la negociación y el acuerdo sobre contratos? ¿Cuántos contratos contienen términos que establecen cómo se tiene que utilizar la información sobre efectividad?
6. *Incentivos:* ¿qué incentivos, tanto individuales como organizacionales, existen para fomentar la práctica de la medicina basada en la evidencia? ¿Qué desincentivos existen para desalentar la práctica inadecuada y las modificaciones injustificadas en la toma de decisiones clínicas?
7. *Sistemas de información:* ¿se está usando al máximo el potencial de los sistemas de información existentes para monitorizar la efectividad clínica?

¿Existe algún argumento comercial para emplear los nuevos sistemas de información con el fin de abordar las tareas, y se está teniendo en cuenta esta cuestión al tomar las decisiones sobre compra de tecnología de la información?

8. *Auditoría clínica:* ¿hay un programa real de auditoría clínica en toda la organización, capaz de abordar las cuestiones sobre la efectividad clínica y lograr los cambios adecuados en la práctica?

Apéndice 2 Evaluación de los efectos de una intervención

	Evento		*Total*
	Sí	*No*	
Grupo control	*a*	*b*	*a + b*
Grupo experimental	*c*	*d*	*c + d*

Si el evento no es deseable (p. ej., fallecimiento)
CER = riesgo de un resultado no deseable en el grupo control = *a/(a + b)*
EER = riesgo de un resultado no deseable en el grupo experimental = *c/(c + d)*
Riesgo relativo del evento no deseable en el grupo experimental frente al control = EER/CER
Reducción absoluta del riesgo en el grupo tratado (ARR) = CER – EER
Número necesario a tratar (NNT) = 1/ARR = 1/(CER – EER)

Si el evento es deseable (p. ej., curación)
CER = riesgo de un resultado deseable en el grupo control = *a/(a + b)*
EER = riesgo de un resultado deseable en grupo experimental = *c/(c + d)*
Aumento relativo del beneficio en el grupo tratado frente al control = EER/CER
Aumento absoluto del beneficio en el grupo tratado frente al control = EER – CER
Número necesario a tratar (NNT) = 1/ARR = 1/(EER – CER)

Agradecimientos

Quiero agradecer a Paul Glasziou del Oxford Centre for Evidence-Based Medicine sus aclaraciones sobre estos conceptos.