CUADERNOS DE LA FUNDACIÓN DR. ANTONIO ESTEVE **Nº26**

Bioestadística para periodistas y comunicadores

Coordinador: Gonzalo Casino

FUNDACIÓN DR. ANTONIO ESTEVE

AECC
Asociación Española de Comunicación Científica

CUADERNOS DE LA FUNDACIÓN DR. ANTONIO ESTEVE **Nº 26**

Bioestadística para periodistas y comunicadores

Coordinador: Gonzalo Casino

FUNDACIÓN DR. ANTONIO ESTEVE

AECC
Asociación Española de Comunicación Científica

La presente edición recoge la opinión de sus autores, por lo que la Fundación Dr. Antonio Esteve no se hace necesariamente partícipe de su contenido.

© 2013, Fundación Dr. Antonio Esteve
Llobet i Vall-Llosera 2. E-08032 Barcelona
Teléfono: 93 433 53 20; fax: 93 450 48 99
Dirección electrónica: fundacion@esteve.org
http://www.esteve.org

Depósito Legal: GI. 495-2013
ISBN: 978-84-940656-5-1

La Fundación Dr. Antonio Esteve, establecida en 1983, contempla como objetivo prioritario el estímulo del progreso de la farmacoterapéutica por medio de la comunicación y la discusión científica.

La Fundación quiere promover la cooperación internacional en la investigación farmacoterapéutica y, a tal fin, organiza reuniones internacionales multidisciplinarias donde grupos reducidos de investigadores discuten los resultados de sus trabajos. Estas discusiones se recogen diferentes formatos de publicación como los *Esteve Foundation Symposia* y los *Esteve Foundation Discussion Groups.*

Otras actividades de la Fundación Dr. Antonio Esteve incluyen la organización de reuniones dedicadas a la discusión de problemas de alcance más local y publicadas en formato de monografías o cuadernos. La Fundación participa también en conferencias, seminarios, cursos y otras formas de apoyo a las ciencias médicas, farmacéuticas y biológicas, entre las que cabe citar el Premio de Investigación que se concede, con carácter bienal, al mejor artículo publicado por un autor español dentro del área de la farmacoterapia.

Entre la variedad de publicaciones que promueve la Fundación Dr. Antonio Esteve, cabe destacar la serie *Pharmacotherapy Revisited* en la cual a través de diferentes volúmenes se recopilan, en edición facsímil, los principales artículos que sentaron las bases de una determinada disciplina.

Índice

Presentación
Gonzalo Casino VII

Participantes IX

Los periodistas ante la bioestadística: problemas, errores y cautelas
Gonzalo Casino 1

¿Qué pretende y qué puede contestar la investigación biomédica?
Erik Cobo 11

La epidemiología y los estudios observacionales de cohortes y de casos y controles
José Luis Peñalvo 19

La confianza en los resultados de la investigación y el sistema GRADE
Pablo Alonso 25

Diálogo 1. Herramientas estadísticas, buenos consejos y cierta intuición periodística
Ainhoa Iriberri 33

Diálogo 2. Sobre los estudios observacionales y su tratamiento periodístico
Pablo Francescutti 39

Talleres. Análisis de *papers,* comunicados de prensa y artículos periodísticos
Esperanza García Molina 47

Seeing through the hype: problems with media coverage and how to do better
Steven Woloshin and Lisa M. Schwartz 55

Seeing through the hype: Garbage! When the news is not fit to print
Lisa M. Schwartz and Steven Woloshin 63

33 mensajes clave 69

Bibliografía recomendada 73

Apéndice
Numbers glossary 77
Statistics glossary 78
Questions to guide your reporting 79
How to highlight study cautions 80

Presentación

Gonzalo Casino

Esta publicación quiere dar cuenta de la jornada sobre bioestadística para periodistas y comunicadores organizada al alimón por la Asociación Española de Comunicación Científica (AECC) y la Fundación Dr. Antonio Esteve, que se celebró en Madrid el 14 de febrero de 2013, año internacional de la estadística. Además de cumplir este objetivo, pretende servir de guía y orientación para los informadores que se enfrentan a un estudio biomédico. Para ello, compensa la ausencia de una exposición formal de los conceptos estadísticos básicos con numerosas indicaciones y pautas para mejorar la información sobre los resultados de la investigación médica.

Informar sobre la investigación de los problemas de salud y sobre los riesgos y beneficios de las intervenciones médicas es un asunto complicado para los periodistas, entre otras cosas porque la mayoría no ha recibido formación para interpretar estadísticas de salud. (Digamos, entre paréntesis, que muchos médicos tampoco las saben interpretar correctamente.) Casi todos tenemos alguna dificultad para entender las estadísticas –y en consecuencia para tomar decisiones sobre nuestra salud– porque en la escuela no se enseñan las matemáticas de la incertidumbre sino las de la certeza. Sin embargo, en el ámbito de la salud apenas hay certezas, todo son probabilidades, y una de las funciones básicas de los periodistas que informan sobre biomedicina es precisamente explicar esta incertidumbre.

La formación en bioestadística es una de las grandes carencias de los periodistas y comunicadores científicos. Cuando en nombre de la AECC propuse a Fèlix Bosch, director de la Fundación Dr. Antonio Esteve, coorganizar una jornada de estas características, enseguida quedó claro que, por los intereses y antecedentes de ambas entidades, éramos los socios idóneos para esta iniciativa. Desde un primer momento convinimos perfilar una jornada formativa que combinara la teoría y la práctica, con el máximo rigor y lo más pegada posible al quehacer cotidiano de los informadores, con presentaciones breves y talleres prácticos con casos reales, complementado todo ello con diálogos entre periodistas y estadísticos. Y que todo esto quedara plasmado en una publicación que pudiera servir de guía y referencia para periodistas y comunicadores.

Encapsular por completo estos contenidos teóricos y prácticos en el limitado margen de una jornada hubiera sido imposible sin la estrecha colaboración y el entusiasmo de Erik Cobo y Pablo Alonso, auténticos cómplices en el diseño del programa y protagonistas incansables de los diálogos y talleres. La presencia en la jornada –y en este libro– de Lisa M. Schwartz y Steven Woloshin, expertos entre los expertos en comunicación de resultados biomédicos, es una garantía de máxima calidad. Cuando aceptaron participar, supe que los periodistas no se irían con las manos vacías. Sin miedo a exagerar, se puede afirmar que esta pareja de médicos ha liderado las principales investigaciones sobre la adecuación a las pruebas científicas de las informaciones periodísticas, los comunicados de prensa y la publicidad. Las abundantes investigaciones de estos maestros del rigor y la ponderación en los mensajes de salud, así como los textos incluidos en este libro, hablan por sí solos.

Mis colegas periodistas Ainhoa Iriberri, Esperanza García Molina y Pablo Francescutti, com-

pañeros en los diálogos, dieron buenas muestras de que el periodismo científico es probablemente más necesario que nunca, como también atestiguaron algunos de los asistentes. Lo acontecido en las casi 10 horas de jornada desborda los límites de esta publicación, cuyo éxito se debió en buena medida al apoyo organizativo de Gonzalo Remiro y Patricia Medrano, por parte de la AECC, y de Pol Morales por parte de la Fundación Dr. Antonio Esteve, así como a la colaboración del Centro Nacional de Investigaciones Cardiovasculares (CNIC), que acogió la jornada y estuvo representado por el epidemiólogo José Luis Peñalvo. Su director, Valentín Fuster, nos dio la bienvenida; Fèlix Bosch inauguró la jornada, y Antonio Calvo, presidente de la AECC, la cerró con unas reflexiones sobre el oficio de informar.

Entre todos los ponentes hemos elaborado una lista de 33 mensajes clave, que figura al final del libro como colofón, guía e invitación para mejorar la práctica del oficio de informador. Como complemento, el ilustrador Enrique Ventura ha creado una serie de viñetas que aporta un contrapunto lúdico y escéptico para reflexionar sobre la interpretación de las estadísticas en biomedicina.

Para que la muestra sea representativa hace falta azar

Behar / Ventura

La idea de incluir una serie de viñetas de humor la planteó Erik Cobo, en una conversación con Pablo Alonso, Fèlix Bosch y Gonzalo Casino. Enrique Ventura sumó con entusiasmo su profesionalidad y todos recogimos ideas del entorno. Nuestro especial agradecimiento a José Antonio González, Matt Elmore y Jordi Cortés por sus ideas y críticas. Y a Roberto Behar, Pera Grima y Lluís Marco por ser fuente de inspiración (Behar R, Grima P, Marco-Almagro L. *Twenty-five analogies for explaining statistical concepts.* JASA. 2013;67:44-48).

Participantes

Pablo Alonso Coello
Especialista en Medicina Familiar y Comunitaria e investigador Miguel Servet del Centro Cochrane Iberoamericano (Instituto de Investigación Biomédica Sant Pau), en Barcelona. Su investigación se centra en la metodología de las guías de práctica clínica, revisiones sistemáticas y ensayos clínicos, y en la toma de decisiones por parte de los pacientes.

Gonzalo Casino
Periodista científico y licenciado en Medicina con posgrado en Bioestadística. Coordinador de la información de salud de *El País* durante una década y director editorial de Doyma/Elsevier España. Actualmente dirige la revista *Técnica Industrial* y escribe para *El País, Intramed* y *The Lancet.*

Erik Cobo
Estudió Medicina en Barcelona y Estadística en Essex y París. Es profesor titular del Departamento de Estadística e Investigación Operativa de la UPC de Barcelona, editor de metodología de *Medicina Clínica* y editor asociado de *Trials.* Uno de sus últimos libros es *Bioestadística para no estadísticos.*

Pablo Francescutti
Periodista científico, profesor de la Facultad de Ciencias de la Comunicación de la Universidad Rey Juan Carlos de Madrid y secretario de la Asociación Española de Comunicación Científica. Autor de varios estudios sobre comunicación científica.

Esperanza García Molina
Redactora jefa de la agencia SINC y vicepresidenta de la Asociación Española de Comunicación Científica. Es licenciada en Física por la UCM y máster en Periodismo y Comunicación de la Ciencia, la Tecnología y el Medio Ambiente por la UC3M.

Ainhoa Iriberri Moreno
Periodista especializada en biomedicina. Empezó su carrera en *El Mundo* en 2000 y ha trabajado también en *Reuters* y *Público.* Actualmente es periodista *freelance* para varios medios, como *Muy Interesante, British Medical Journal* y SINC, entre otros.

José Luis Peñalvo
Licenciado en Farmacia y doctor en Nutrición, es investigador del Departamento de Epidemiología y Genética de Poblaciones del Centro Nacional de Investigaciones Cardiovasculares. Su investigación se centra en el estudio de los factores de riesgo cardiovascular, en colaboración con la Johns Hopkins Bloomberg School of Public Health, de Estados Unidos.

Lisa M. Schwartz y Steven Woloshin
Profesores de Medicina Familiar y Comunitaria en la Geisel School of Medicine, en Dartmouth (Estados Unidos), y codirectores del Center for Medicine and the Media en el Dartmouth Institute for Health Policy and Clinical Practice. Investigan las esperanzas y temores desproporcionados creados por las revistas médicas, la publicidad y los medios de comunicación. Participan habitualmente en cursos de formación para periodistas en Estados Unidos. Sus estudios se han publicado en las principales revistas médicas, y han escrito en diarios como *The New York Times* y el *Washington Post*.

No todas las preguntas de salud necesitan responderse con un ensayo clínico

Alonso / Ventura

Los periodistas ante la bioestadística: problemas, errores y cautelas

Gonzalo Casino

Detrás de cada mensaje de salud hay –o debería haber– números. Los resultados de la investigación médica se expresan con números y datos estadísticos que los resumen y no son fáciles de entender. La omnipresencia del cálculo de probabilidades en la investigación clínica y epidemiológica hace que la información biomédica sea algo demasiado técnico no sólo para la ciudadanía sino también para los periodistas. Para mejorar la información médica es imprescindible entender las estadísticas de salud. Y para ello conviene analizar el problema, revisar algunos errores habituales y tener presentes ciertas cautelas a la hora de informar. Éste es el propósito de este capítulo.

Un doble problema

Más de tres cuartas partes de los periodistas reconocen tener dificultades con las estadísticas de salud.[1] Este problema lo podemos enunciar de muy diversas maneras: «los *papers* resultan muy difíciles de entender», «la jerga estadística nos desborda», «el periodismo se lleva muy mal con la incertidumbre», «los expertos reconocen que muchas estadísticas se cocinan», etcétera. Pero estas y otras expresiones vienen a decir que a los periodistas científicos nos falta formación estadística. Así pues, tenemos un problema, o más bien un doble problema: no entendemos bien las estadísticas de salud y, por tanto, tenemos dificultades para informar correctamente.

Las causas de este problema son diversas. En primer lugar, la ciencia médica y su herramienta, la bioestadística, son cada vez más sofisticadas; en segundo lugar, la formación de los periodistas no siempre es suficiente, y en tercer lugar, la escasez de voluntad divulgadora entre los investigadores y, a veces, de la necesaria transparen-

cia. La consecuencia de todo ello es que, demasiado a menudo, lejos de informar contribuimos a la desinformación, con el consiguiente posible perjuicio para la salud que esto acarrea.

Bien es verdad que las dificultades con las estadísticas de salud no afectan sólo a los periodistas. Muchos médicos también tienen dificultades. Un reciente estudio de Odette Wegwarth[2] realizado con médicos estadounidenses indica que los clínicos están muy lejos de comprender las estadísticas del cribado. La mayoría de ellos no distingue la información relevante (reducción de la mortalidad) de la no relevante (tasa de supervivencia), se dejan confundir por el engañoso concepto de supervivencia en el cribado, ignoran la influencia del sesgo de anticipación diagnóstica y demuestran una preocupante falta de conocimientos estadísticos básicos.

Así pues, un primer mensaje para los periodistas es que conviene consultar fuentes competentes en estadística.

«Según un estudio»

Nada parece respaldar tanto la veracidad de un mensaje como el aval de un estudio. La muletilla «según un estudio» es moneda corriente en las informaciones periodísticas de salud. La palabra «estudio» tiene las espaldas tan anchas y tan amplias las tragaderas que lo mismo sirve para designar una encuesta de medio pelo que una rigurosa investigación científica, un intrascendente análisis estadístico que un ensayo clínico riguroso. Pero lo cierto es que aludir vagamente a «un estudio» no dice nada si no se añaden a continuación sus datos esenciales. De esta imprecisión y calculada ambigüedad se aprovechan, obviamente, los trabajos más chapuceros, que no sólo se utilizan para publicitar los supuestos beneficios de un producto o una intervención sino que a veces encuentran eco en informaciones periodísticas.

La pirámide de la evidencia

Los estudios de salud pueden clasificarse de diferentes maneras en función de su diseño. Una primera y elemental clasificación de la investigación médica con personas distingue entre estudios observacionales y experimentales o de intervención, es decir, entre los que se limitan a observar las características de una población y los que realizan una intervención sobre ella (tratamiento, prueba diagnóstica, etcétera). Estos últimos son básicamente los ensayos clínicos, mientras que los estudios observacionales pueden ser de distinto tipo, aunque los más habituales son las series de casos, los estudios transversales, los estudios de casos y controles, y los estudios de cohortes.

Aparte de los estudios de intervención y de los observacionales, realizados con personas, pueden considerarse también los estudios in vitro y los estudios preclínicos o con animales, que son un primer escalón en las investigaciones de salud y que a menudo suscitan también interés informativo. Unos y otros conforman la denominada investigación primaria, mientras que la secundaria sería la realizada a partir de ésta. Entre los estudios secundarios, los de mayor interés informativo son las revisiones sistemáticas, con o sin metaanálisis.

La confianza que merecen los resultados de todos estos tipos de estudios es muy distinta. En función del diseño, clásicamente se han jerarquizado en una pirámide, la llamada «pirámide de la evidencia», o mejor dicho de las pruebas científicas (fig. 1). En esta pirámide se observa un

Figura 1. La pirámide de la evidencia (pruebas científicas) jerarquiza los distintos tipos de estudios según la confianza que a priori merecen por su diseño.

gradiente de calidad o confianza exclusivamente según el tipo de estudio. No obstante, el diseño no lo es todo y es necesario considerar otros aspectos, como la consistencia o la precisión de los resultados, para conocer la confianza que éstos merecen (véase el capítulo *La confianza en los resultados de la investigación y el sistema GRADE*).

Así pues, hay que tener presente que referirse en un artículo periodístico a «un estudio» es demasiado vago, porque no todos los estudios son iguales ni merecen la misma confianza. Es imprescindible informar de sus características y de la confianza que merecen los resultados.

Algunos errores habituales

Aunque muchas piezas periodísticas son impecables, realmente no es difícil encontrar errores en las informaciones sobre estudios de salud. En descargo de los periodistas, hay que hacer notar que muchos de estos errores ya vienen inducidos por las fuentes, los comunicados de prensa y otros intermediarios de la información biomédica. Por su importancia y recurrencia, voy a llamar la atención sobre tres errores habituales.

Enfatizar el riesgo relativo y olvidar el riesgo absoluto

Una de las formas habituales de expresar los resultados de las investigaciones de salud es en forma de riesgos. Y uno de los principales errores que se cometen en la comunicación de riesgos es omitir los valores absolutos e indicar sólo los relativos, que muestran de forma muy expresiva la magnitud de un efecto en los ensayos clínicos o la asociación entre una exposición y un efecto en los estudios observacionales.[3] Decir, por ejemplo, que un fármaco reduce un 50% el riesgo de muerte o que una exposición lo aumenta un 100% puede resultar muy impactante, pero también puede ser engañoso si estos valores no se acompañan de los correspondientes valores absolutos. No es lo mismo reducir el riesgo de muerte del 10% al 5% que hacerlo del 0,2% al 0,1% (50% en ambos casos); tampoco es lo mismo aumentarlo del 10% al 20% o del 0,1% al 0,2% (100%, en términos relativos, en ambos casos). Y es que si un riesgo es extremadamente bajo, aunque se reduzca a la mitad o se duplique seguirá siendo muy bajo.

Como consecuencia de esta deficiente información, el público puede tener una percepción del riesgo equivocada. Culpar de esta situación a los periodistas es fácil, pero el problema es más complejo. La mayoría de los artículos publicados en las principales revistas biomédicas no indica en los resúmenes los riesgos absolutos, y la mitad de esos artículos ni siquiera los mencionan en el resto del texto. Un análisis de los artículos científicos sobre mediciones de riesgos para la salud publicados durante un año en las principales revistas de medicina mostró que en muchos de ellos no figura el valor de riesgo absoluto, un dato esencial para valorar en su justa medida un riesgo para la salud.[4] Además, en los *press releases* que distribuyen las revistas biomédicas, y que suelen ser el punto de partida para la elaboración de las informaciones periodísticas, tampoco suelen aparecer.[5]

Con todo, los periodistas no debemos enfatizar un riesgo relativo olvidando el riesgo absoluto, que es el que mejor ilustra la dimensión de un problema de salud.

Mitificar la prevención

La idea de que es mejor prevenir que curar goza de tal prestigio y difusión que cualquier argumentación en contra parece poco menos que un desvarío. En medicina, las exploraciones colectivas o cribados *(screening)* de ciertas enfermedades se ven con general aprobación, sin reparar en que estas pruebas, aparte de un coste importante, tienen también sus riesgos. La idea de que la detección precoz no siempre es la mejor opción resulta difícil de cuestionar, pues es contraintuitiva y sólo es posible llegar a ella tras una rigurosa ponderación de los riesgos y los beneficios.

Los mensajes que defienden el cribado, avalados por médicos y autoridades sanitarias, están por todas partes y a veces incluso respaldados con la imagen y el testimonio de famosos. ¿Cómo vamos a ponerlos en duda? ¿Acaso la mamografía no ayuda al diagnóstico precoz del

Tabla 1. Beneficios y riesgos del cribado del cáncer de mama con mamografía.

	Beneficio: reducción del riesgo de muerte (10 años)	
	Edad: 40-49 años	Edad: 50-59 años
Sin cribado	3,5/1000	5,3/1000
Con cribado	3,0/1000	0,7/1000
	Perjuicios: angustia, biopsias, sobretratamientos	
	Edad: 40-49 años	Edad: 50-59 años
Falsos positivos + biopsia	60-200/1000	50-200/1000
Sobrediagnósticos	1-5/1000	1-7/1000

cáncer de mama y a evitar sufrimientos en muchas mujeres? ¿Acaso la prueba del PSA (antígeno prostático) no ayuda a detectar el cáncer de próstata y a reducir su mortalidad? Sin embargo, algunos análisis y artículos en las principales revistas médicas han puesto de manifiesto una tendencia a sobrestimar los beneficios del cribado y subestimar sus riesgos.

Steven Woloshin y Lisa M. Schwartz ilustran con números sencillos los riesgos y beneficios del cribado del cáncer de mama con mamografía.[6] Para las mujeres de 50 a 59 años de edad, el beneficio del cribado se resume en reducir el riesgo de muerte a 10 años de 5,3 a 4,6 mujeres por cada 1000 revisadas anualmente durante 10 años, es decir, apenas se evita la muerte de una de cada 1000; el riesgo del cribado se cifra en que de 50 a 200 de cada 1000 serán sometidas innecesariamente a una biopsia por un falso positivo, y entre 1 y 7 de cada 1000 serán tratadas por un cáncer que no tienen. Y los números para las mujeres de 40 a 49 años de edad son todavía más elocuentes (tabla 1).

La prevención tiene, por tanto, un precio no sólo económico. Los falsos positivos y los tratamientos innecesarios representan mucho sufrimiento inútil. Por cada persona que podrá sobrevivir al cáncer gracias a la detección precoz hay otras muchas que serán sometidas a pruebas y tratamientos innecesarios por un cáncer que no tienen. Conocer estos números, presentados de forma clara y con sus riesgos absolutos, es el primer paso para sopesar los riesgos y beneficios y tomar una decisión. Por desgracia, como dicen Woloshin y Schwartz, promover las decisiones informadas es más difícil que vender el cribado.

El psicólogo Gerd Gigerenzer ha realizado un estudio revelador sobre la percepción de los beneficios del cribado de los cánceres de mama y próstata en Europa.[7] Los resultados muestran que el 92% de las mujeres de nueve países europeos, entre ellos España, sobrevalora o ignora

Tabla 2. La población europea sobrevalora el beneficio de las mamografías en la curación del cáncer.

	Porcentaje de respuestas			
Reducción riesgo muerte (x1000 en 10 años)	9 países UE (n=10.288)	España	Alemania	Francia
Ninguna	6,4	3,9	1,4	0,8
1	1,5	2,7	0,8	1,3
10	11,7	6,9	12,8	15,7
50	18,9	11,7	21,3	21,7
100	15,0	11,3	16,8	21,5
200	15,2	15,7	13,7	23,7
No sabe	31,2	48,0	33,1	15,3

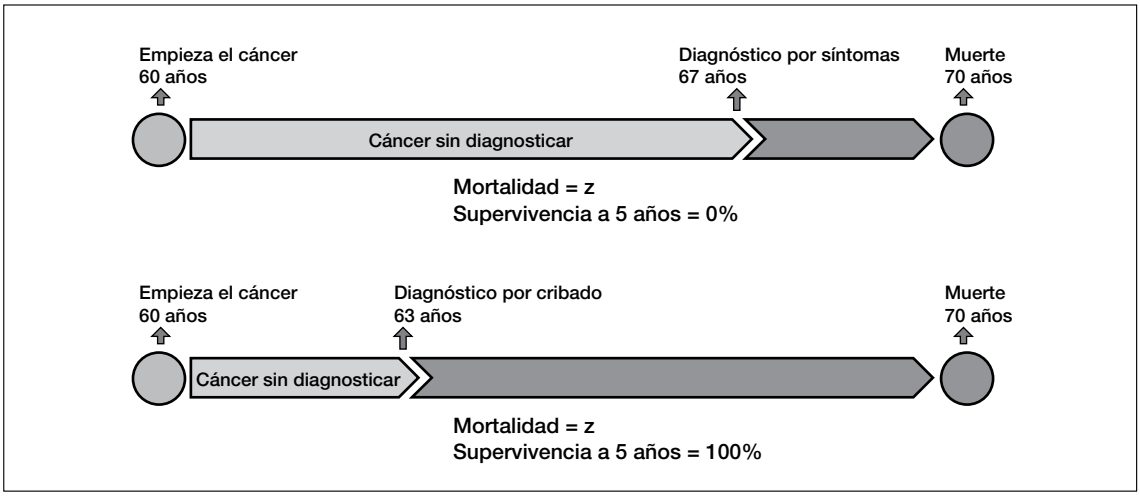


Figura 2. Una mayor tasa de supervivencia no implica necesariamente una menor mortalidad, porque hay que considerar el sesgo por anticipación en el diagnóstico del cáncer.

el efecto real de las mamografías en la reducción de la mortalidad por cáncer de mama (muchas creen que las mamografías salvan vidas en una proporción 10, 50, 100 o incluso 200 veces superior a la real). Asimismo, el 89% de los hombres europeos sobrevalora o ignora el efecto del cribado del cáncer de próstata mediante la determinación del PSA. Lo que revela el trabajo de Gigerenzer, un experto en comunicación de riesgos, es que la población no está bien informada para poder tomar decisiones sobre su salud (tabla 2).

Considerar que una mayor tasa de supervivencia implica más tiempo de vida

El beneficio del cribado suele comunicarse en forma de tasas de supervivencia, que pueden ser muy llamativas pero que no son una prueba del alargamiento del tiempo de vida ni, por tanto, del beneficio del cribado. La supervivencia, tal como se define en el cribado del cáncer, es un indicador del número de personas que tras ser diagnosticadas siguen vivas al cabo de un tiempo, generalmente 5 o 10 años. Pero el tiempo que media entre el diagnóstico y la muerte (supervivencia) depende mucho del momento del diagnóstico.

Imaginemos, por ejemplo, un grupo de pacientes a quienes se diagnostica un cáncer a los 67 años de edad y que acaban muriendo a los 70 años; como sobreviven sólo 3 años, la tasa de supervivencia (a 5 años) es del 0%. Ahora bien, supongamos que ese mismo grupo se somete a un cribado a los 60 años de edad, que a todos ellos se les diagnostica un cáncer y que mueren también a los 70 años; como sobreviven 10 años, la tasa de supervivencia (igualmente a 5 años) es del 100%. Lo que ilustra este escenario hipotético que se explica en el citado artículo de Odette Wegwarth[1] es que la tasa de supervivencia, a pesar de su impresionante aumento de 0 a 100, no demuestra el beneficio del cribado, ya que no hay reducción de la mortalidad (fig. 2).

No se puede, por tanto, comparar la supervivencia entre los diagnosticados (anticipadamente) mediante una prueba de cribado (una mamografía, por ejemplo) y los que han sido diagnosticados cuando se presentan los síntomas del cáncer. Adelantar el diagnóstico implica aumentar el tiempo de conocimiento de la enfermedad, pero no por ello el tiempo de vida.

Las estadísticas de supervivencia se ven distorsionadas asimismo por el diagnóstico precoz de casos de cáncer que no progresan (por ejemplo, una gran proporción de los carcinomas ductales in situ de la mama) y que por tanto nunca darían síntomas. El cribado los saca a la luz y los contabiliza como casos de cáncer, inflando así

las estadísticas de supervivencia. Al comparar la supervivencia del grupo sometido a cribado con la del grupo control, aparece una tasa muy elocuente, aunque la reducción de la mortalidad no lo sea tanto. Ignorar que existe este sesgo por anticipación del diagnóstico *(lead-time bias)* e interpretar erróneamente las estadísticas de supervivencia hace que muchos médicos tengan un entusiasmo desmedido por el cribado.

Detectar más casos de cáncer no quiere decir, por tanto, que una prueba de cribado sea beneficiosa. La prevención y el diagnóstico precoz también tienen sus riesgos, en el caso del cribado, en forma de falsos diagnósticos, sobrediagnósticos y sobretratamientos. Éstas son algunas de las ideas que chocan con la sabiduría popular y con el conocimiento médico convencional. Pero ser un buen periodista científico implica cuestionarse ciertos prejuicios y, también, tener presentes estas cuestiones.

Así pues, conviene recordar que las tasas de supervivencia del cribado pueden crear confusión, y que la reducción de la mortalidad sólo se demuestra con un ensayo clínico que compare la mortalidad de un grupo cribado con otro que no lo ha sido.

Cautelas a la hora de informar

Cuando se maneja un material tan sensible como las estadísticas de salud, todas las cautelas son pocas a la hora de informar. A continuación quiero llamar la atención sobre algunas de las más importantes.

Cuidado con la información preliminar

La divulgación en los medios de comunicación de resultados de investigaciones preliminares, es decir, antes de que hayan concluido y de ser revisadas por expertos, como es usual en la comunidad científica, plantea un problema importante.

La cobertura mediática de estos resultados preliminares, en general presentados en congresos y otros eventos, puede trasladar al público la falsa impresión de que los datos presentados están maduros y son consistentes, que la metodología empleada es fiable y contrastada, y que los resultados de la investigación están ampliamente aceptados, cuando esto no suele ser así, porque las investigaciones comunicadas en los congresos suelen estar en sus etapas iniciales. Lo cierto es que muchas de estas comunicaciones tienen un diseño imperfecto, se basan en muestras pequeñas o en estudios de laboratorio o con animales. Además, el 25% de los trabajos preliminares que han recibido atención mediática permanecen sin publicar en la literatura médica después de 3 años desde su presentación en un congreso.[8]

A pesar de estos peligros, la difusión de resultados preliminares está muy extendida. Un estudio[9] realizado sobre más de 50 periódicos y revistas de gran circulación publicados en inglés ha mostrado que sólo el 57% de las noticias de biomedicina que saltan a primera página de los periódicos están basadas en investigaciones revisadas por expertos y publicadas en revistas revisadas por pares, mientras que la cuarta parte (24%) de las informaciones periodísticas se basan en investigaciones preliminares que siguen sin publicarse en revistas de prestigio 3 años después de aparecer en un periódico.

Como indican Steven Woloshin y Lisa M. Schwartz, «mucho del trabajo presentado en congresos no está listo para consumo público».[10] A la vista del carácter preliminar de los resultados difundidos en congresos, los periodistas que informan de biomedicina deben plantearse la cobertura de estos eventos médicos y valorar, en cada caso, hasta qué punto interesa la información preliminar a la ciudadanía, sobre todo si no se hacen las oportunas advertencias sobre sus limitaciones y contextualización.

Estar alerta ante las posibles exageraciones de los comunicados de prensa

Los estudios realizados sobre la calidad de los comunicados de prensa *(press releases)* de biomedicina indican que, en general, distan mucho de reflejar objetivamente los resultados de la investigación que tratan de divulgar e interpretar. Los sesgos y otras deficiencias observadas están presentes no sólo en los comunicados de prensa elaborados por la industria farmacéutica,

sino también en los que proceden de centros universitarios y de las propias revistas médicas.[4,11-13]

¡Localizar el paper *y leer el resumen!*

Para informar correctamente acerca de una investigación biomédica publicada en una revista científica no basta con tener información indirecta de esta publicación, a través de un comunicado de prensa o de otra fuente. Conviene localizar el estudio original mediante el enlace que suelen ofrecer los comunicados de prensa o con una búsqueda en Internet. En la base de datos de PubMed (http://www.ncbi.nlm.nih.gov/pubmed) están disponibles al menos los resúmenes de muchos de estos artículos.

Los papers *pueden tener errores estadísticos*

Incluso los mejores trabajos de investigación biomédica, que aparecen publicados en las revistas de más prestigio, pueden contener errores. Un análisis realizado por Emili García-Berthou[14] reveló que uno de cada cuatro estudios incluidos en *British Medical Journal*, una de las cinco grandes revistas médicas, tiene errores en los datos estadísticos, mientras que en la otra revista analizada, la reputada *Nature*, el 38% de los artículos incluye algún error.

La revisión de los cálculos estadísticos de los trabajos publicados en estas dos revistas británicas durante un año reveló incoherencias en el 11% de los resultados. Los errores más frecuentes son de redondeo de los números y de transcripción de los resultados. Pero si ya se detectan errores en estos datos verificables, según García-Berthou, ¿qué habrá en los que no son fácilmente comprobables? Este trabajo muestra, por un lado, que incluso las mejores revistas tienen margen de mejora, y por otro, que no hay que descartar que en los trabajos científicos pueda haber errores estadísticos.

Alerta ante las posibles exageraciones en las estadísticas de prevalencia

Exagerar las estadísticas de prevalencia de las enfermedades es una las estrategias medicalizadoras mejor conocidas. Si una enfermedad tiene más afectados, parece que el problema es mayor y las soluciones médicas más necesarias y acuciantes. Aunque muchos investigadores no se atreverían a decir abiertamente que algunas estadísticas de prevalencia están infladas, algunos sí reconocen que dentro de estos números se incluyen personas con formas muy benignas de la enfermedad en cuestión.

En cualquier caso, si uno se entretiene en realizar una suma de urgencia de las estadísticas de afectados por algunas de las enfermedades que más habitualmente salen en los medios de comunicación, se sorprendería de la cantidad de dolencias per cápita que encuentra. Ante esta situación, es responsabilidad del periodista contrastar todas las estadísticas de prevalencia que introduzca en una información, y sospechar que los datos pueden estar inflados.

Parece de Perogrullo, pero los animales no son humanos

Los resultados de la investigación realizada con ratas y otros animales de laboratorio no siempre pueden generalizarse a otros mamíferos, y mucho menos extrapolarse directamente a los seres humanos. Sin embargo, en los medios de comunicación este salto en el vacío se da a menudo. Y una vez más, esta situación no sólo es responsabilidad del periodista. Los elementos previos de la cadena de producción también favorecen la información errónea y sensacionalista. Como caricaturizaba el oncólogo Josep Baselga en una entrevista, «un investigador básico ve una célula y se cree que es un paciente».[15]

Vigilar la terminología

El periodista escribe para ser leído y entendido. Todo su oficio se orienta en esta dirección y, así, es imprescindible manejar con buen juicio y mesura la jerga científica. El mensaje ha de ser entendido por una persona no experta, y por ello es necesario que el periodista comprenda bien aquello de lo que está hablando.

Como advierte el veterano periodista científico Tim Radford,[16] «no escribes para impresionar

al científico a quien acabas de entrevistar, ni al profesor que fue decisivo para tu graduación, ni al editor estúpido que te rechazó o a esa persona tan atractiva que acabas de conocer en la fiesta y sabía que eras periodista (o a su madre). Escribes para impresionar a alguien que está colgado de la barra del metro (…) y que dejaría de leerte en un milisegundo si pudiera hacer algo mejor». Por ello, añade: «Cuidado con las palabras largas y absurdas, con la jerga. Esto es doblemente importante si eres periodista científico, pues de vez en cuando tendrás que manejar palabras que no utiliza ningún ser humano normal: fenotipo, mitocondria, inflación cósmica, campana de Gauss, isostasia…».

Significaciones que no significan nada

Muchos investigadores médicos parecen creer que si no encuentran algo «estadísticamente significativo» no hay nada que valga la pena mostrar. O dicho al revés: basta encontrar algo «estadísticamente significativo» para que el trabajo merezca ser publicado y tenido en cuenta, porque esa significación estadística es un marchamo de calidad. Con esta actitud, nefasta y engañosa como pocas en la investigación médica, lo que se ha conseguido es inundar la literatura de significaciones que no significan nada en la práctica médica y, a la postre, que estos resultados tengan eco en los medios de comunicación. ¿Tiene acaso relevancia que un fármaco contra el cáncer pueda alargar «significativamente» la vida del enfermo durante un mes a cambio de una peor calidad de vida?

En 2008, la revista *Nature* llamaba la atención sobre algunos de los términos científicos más difíciles de definir,[17] y uno de los ocho elegidos era precisamente «significativo», un adjetivo que parece ilustrar por sí mismo la importancia de un descubrimiento. Pero esto no es así, entre otras cosas porque el concepto de significación estadística está lejos de ser comprendido por la mayoría de los científicos, según afirma en *Nature* el bioestadístico Steven Goodman. Decir que una asociación entre dos variables es estadísticamente significativa quiere decir que puede descartarse que haya aparecido por azar, porque si no hubiera dicha asociación, resultados como el observado serían muy poco probables (normalmente con una p o probabilidad menor de 0,05 o de 0,01). En el caso de un tratamiento, un resultado significativo ($p < 0,05$) quiere decir que hay una posibilidad menor del 5% de observar (por azar) ese resultado favorable aun cuando el tratamiento no sea eficaz.

Sin embargo, «desde el punto de vista clínico la significación estadística no resuelve todos los interrogantes que hay que responder, ya que la asociación estadísticamente significativa puede no ser clínicamente relevante y, además, la asociación estadísticamente significativa puede no ser causal», como escriben Salvador Pita Fernández y Sonia Pértega Díaz en un esclarecedor artículo,[18] donde añaden que «podemos encontrar asociaciones estadísticamente posibles y conceptualmente estériles».

Y es que una cosa es la significación estadística, otra la significación médica o relevancia clínica, y una tercera la significación periodística o interés mediático, sin el cual difícilmente una investigación tendrá eco en los medios. En medicina, como advierten Pita y Pértega, «cualquier diferencia entre dos variables estudiadas puede ser significativa si se dispone del suficiente número de pacientes». Así pues, la significación estadística puede ser también un camino hacia la irrelevancia y la confusión. Y esto es algo que hay que tener en cuenta a la hora de decidir si se informa de una investigación.

Atención al texto y al contexto

En la información científica, el contexto es fundamental. Sin contextualizar los resultados de una investigación, el estudio en cuestión no deja de ser una anécdota. Podemos entender la investigación como una conversación continuada, como una discusión coral a lo largo del tiempo en la cual unos investigadores replican a otros, se respaldan o se desdicen con sus respectivos estudios. Un estudio sería, pues, como una frase en medio de una conversación, de modo que para entenderla debidamente hay que conocer de qué están hablando los investigadores y qué han dicho.

El periodista debe, por tanto, informar de la conversación, del contexto en que se realiza el

estudio en cuestión. Y para ello no sólo tiene que hablar con los protagonistas del estudio, sino también con fuentes independientes que le ayuden a contextualizar los nuevos resultados.

Bibliografía

1. Voss M. Checking the pulse: Midwestern reporters' opinions on their ability to report health care news. Am J Public Health. 2002;92:1158-60.
2. Wegwarth O, Schwartz LM, Woloshin S, Gaissmaier W, Gigerenzer G. Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. Ann Intern Med. 2012;156:340-9.
3. Casino G. Producers, communicators and consumers of 'risk'. J Epidemiol Community Health. 2010;64:940.
4. Schwartz LM, Woloshin S, Dvorin EL, Welch HG. Ratio measures in leading medical journals: structured review of accessibility of underlying absolute risks. BMJ. 2006;333:1248.
5. Woloshin S, Schwartz LM. Press releases: translating research into news. JAMA. 2002;287:2856-8.
6. Woloshin S, Schwartz LM. The benefits and harms of mammography screening: understanding the trade-offs. JAMA. 2010;303:164-5.
7. Gigerenzer G, Mata J, Frank R. Public knowledge of benefits of breast and prostate cancer screening in Europe. J Natl Cancer Inst. 2009;101:1216-20.
8. Schwartz LM, Woloshin S, Baczek L. Media coverage of scientific meetings: too much, too soon? JAMA. 2002;287:2859-63.
9. Lai WY, Lane T. Characteristics of medical research news reported on front pages of newspapers. PLoS One. 2009;4:e6103.
10. Woloshin S, Schwartz LM. Media reporting on research presented at scientific meetings: more caution needed. Med J Aust. 2006;184:576-80.
11. Kuriya B, Schneid EC, Bell CM. Quality of pharmaceutical industry press releases based on original research. PLoS One. 2008;3:e2828.
12. Woloshin S, Schwartz LM, Casella SL, Kennedy AT, Larson RJ. Press releases by academic medical centers: not so academic? Ann Intern Med. 2009; 150:613-8.
13. Puliyel J, Mathew JL, Priya R. Incomplete reporting of research in press releases: et tu, WHO? Indian J Med Res. 2010;131:588-9.
14. García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. BMC Med Res Methodol. 2004;4:13.
15. Millás JJ. Entrevista a Josep Baselga. El País Semanal, 27 de enero de 2002.
16. Radford T. Manifiesto para periodistas sencillos. (Consultado el 20/01/2013.) Disponible en: http://www.paperpapers.net/2012/05/manifiesto-para-periodistas-sencillos.html
17. Ledford H. Language: disputed definitions. Nature. 2008;455:1023-8.
18. Pita Fernández S, Pértega Díaz S. Significancia estadística y relevancia clínica. Cad Aten Primaria. 2001;8:191-5. Actualizado el 19/09/2001. (Consultado el 20/01/2013.) Disponible en: http://www.fisterra.com/mbe/investiga/signi_estadi/signi_estadi.asp

La ciencia no escribe leyes, propone modelos

Cobo / Ventura

¿Qué pretende y qué puede contestar la investigación biomédica?

Erik Cobo

Introducción

Este capítulo ofrece claves muy generales, pero básicas, para interpretar los estudios empíricos. Incluye principios científicos, metodológicos, clínicos y estadísticos que permitirán al lector situar cada estudio en su contexto. La ciencia progresa gracias al contraste entre ideas y datos. Diferentes preguntas médicas requieren distintos y específicos diseños.

Principios generales

Conjeturas y refutaciones

El método científico propone modelos que representan el entorno y los enfrenta con datos recogidos de forma reproducible: la ciencia establece puentes entre ideas y datos. Esta capacidad de ser observado es fundamental, ya que para poder ser considerado científico un modelo debe poder entrar en conflicto con datos futuros observables. Por ejemplo, «los marcianos existen» es una expresión hoy por hoy infalible, en el sentido de que, como es imposible recorrer todo el universo y demostrar que no existen, no puede entrar en conflicto con datos concebibles.

La ciencia, pues, quiere ser falible. Este contraste empírico implica que los modelos científicos son constantemente abandonados en beneficio de otros que los mejoran y matizan. En consecuencia, no se pretende que sean definitivamente ciertos, pero sí que sean útiles y ofrezcan claves para interpretar, mejorar y disfrutar nuestro entorno. Por ejemplo, las leyes de Newton son falsas: fueron refutadas por Einstein, que las modificó para abarcar también largas distancias. Pero los modelos de Newton se siguen usando para hacer casas ¡que se aguantan!

La ciencia no pretende escribir las leyes del universo, tan sólo modelos que lo reproduzcan.

Una cita célebre de George Box dice «todos los modelos son erróneos, pero algunos son útiles».

Inducción frente a deducción

Tenemos una gran tradición en razonamiento deductivo: partiendo de unos principios que no se discuten, matemáticas, derecho o teología deducen sus consecuencias. Pero la predicción de lo que ocurrirá mañana requiere aplicar el conocimiento más allá. El método científico parte del conocimiento disponible para, primero, deducir consecuencias contrastables, y luego, una vez observadas éstas en unos casos, usar la inferencia estadística para inducir los resultados a una población más amplia.

Exploración y confirmación

Esta posibilidad de enfrentar las ideas con sus consecuencias contrastables divide al proceso científico en dos fases. Al inicio del proceso de investigación y desarrollo (I+D), el análisis exploratorio propone un modelo a partir de los datos. Es lícito torturar los datos hasta que canten, pero debe quedar claro: «nuestros resultados sugieren que…». Al final de la I+D, un análisis confirmatorio preespecificado permite decir «hemos demostrado que…».

Se bromea diciendo que un bioestadístico es un profesional que niega que Colón descubriera América porque no estaba en el protocolo de su viaje. Pero en realidad un bioestadístico pediría a Colón lo mismo que los Reyes Católicos: «Qué interesante. Ande, vuelva y confírmelo». El primer viaje fue una atractiva novedad, pero se necesitaron varios más para abrir una nueva vía comercial.

El mérito de Fleming no radicó en inhibir accidentalmente un cultivo. Su mérito fue conjeturar acertadamente qué pasó y luego replicarlo. John P.A. Ioannidis[1] desarrolla un modelo que muestra que los estudios confirmatorios con resultados positivos tienen una probabilidad de ser ciertos del 85%, que baja al 0,1% en los exploratorios. Leah R. Jager y Jeffrey T. Leek[2] estiman que son ciertos un 84% de los resultados positivos de cinco revistas médicas punteras, que podríamos clasificar como confirmatorias. Y Stephen Senn suele bromear diciendo: «disfrute de sus inesperados resultados significativos… ¡porque no los volverá a ver!». En resumen, un estudio exploratorio aporta ideas nuevas; uno confirmatorio ratifica o descarta ideas previas (fig. 1).

Intervención frente a pronóstico

Los modelos pueden construirse con dos objetivos claramente diferenciados. En primer lugar, por su ambición, tenemos los modelos de intervención, que pretenden cambiar la evolución de los pacientes y requieren una relación de causa-efecto que permitirá, mediante intervenciones en la variable causa, modificar el valor futuro de la

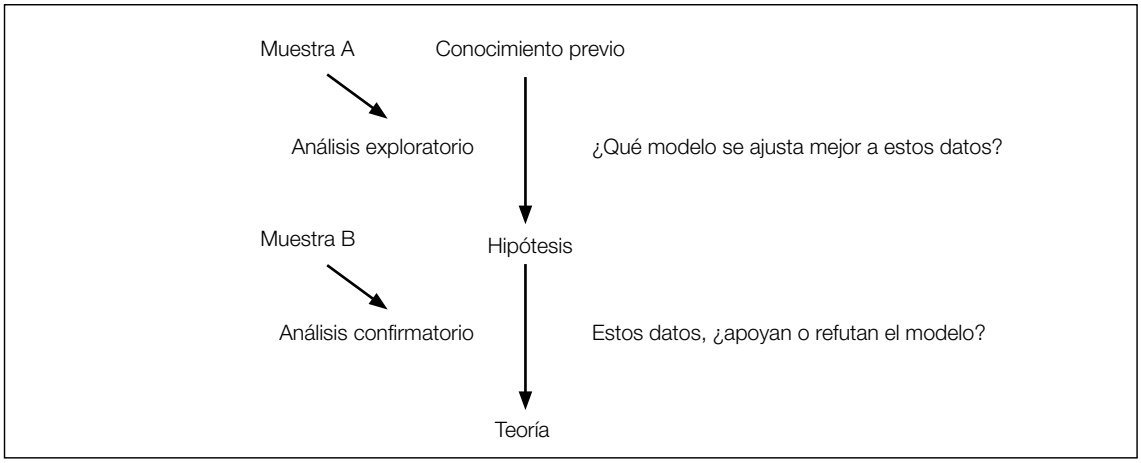


Figura 1. Estudios exploratorios y confirmatorios.

variable respuesta *(outcome*, *endpoint)* o desenlace que sirve para medir el efecto. Por otra parte, tenemos los modelos de relación o predictivos. A diferencia de los anteriores, no precisan una relación de causa-efecto. Se utilizan, por ejemplo, en el diagnóstico y en el pronóstico médico.

Por ejemplo, cuando David me lleva a pescar me pide que observe dónde está agitada el agua en la superficie. Saber que los peces mayores empujan a los menores hacia arriba y que estos baten la superficie, le permite predecir una mayor probabilidad de pesca allí donde el agua está agitada. Usa la agitación como un "chivato". Pero David no sugiere intervenir sobre la agitación del agua para aumentar la probabilidad de pesca.

Otro ejemplo: en la ciudad de Framingham recogieron datos iniciales de una gran cohorte que siguieron muy fielmente durante décadas para observar episodios cardiovasculares. Con la ayuda del modelado estadístico, establecieron grupos con diferente riesgo de presentar un episodio cardiovascular. Entre las variables que contribuían al pronóstico estaba la presión arterial. Una interpretación causal («los que hoy tienen las arterias a reventar, mañana les revientan; ergo, si bajo hoy la presión arterial, bajaré mañana los episodios cardiovasculares») abrió la vía para pensar en intervenciones que bajaran la presión arterial, cuyos efectos fueron estimados en ensayos clínicos. Este ejemplo muestra que un estudio de cohortes cuantifica un pronóstico y lanza interpretaciones causales. En cualquier caso, tanto la intervención como el pronóstico hacen predicciones sobre relaciones que luego deben ser contrastadas.

Medidas del efecto
y medidas de la reducción de la incertidumbre

Para determinar cuánto cambiamos la variable respuesta, recurrimos a medidas de la magnitud del efecto. Dos posibles ejemplos serían: «si toma esta pastilla a diario, bajará 5 mmHg su presión arterial sistólica» o «por cada kilo de peso que pierda, bajará 1 mmHg su presión arterial sistólica».

Para saber cuánto anticipamos de otra variable (presente o futura), recurrimos a medidas de reducción de la incertidumbre. Por ejemplo, «si desconozco la altura de un hombre, mi predicción sobre el peso se centra en su media, 70 kg, con una desviación típica (o error esperado) de 10 kg, pero si conozco que mide 150 cm, mi predicción cambia a 50 kg y la desviación típica alrededor de esta predicción baja a 6 kg»; o también «el peso predice un 15% de la variabilidad de la presión arterial sistólica de un adulto sano».

Tipos de estudios

Objetivos sanitarios

Los objetivos sanitarios se traducen en diferentes preguntas científicas (fig. 2).

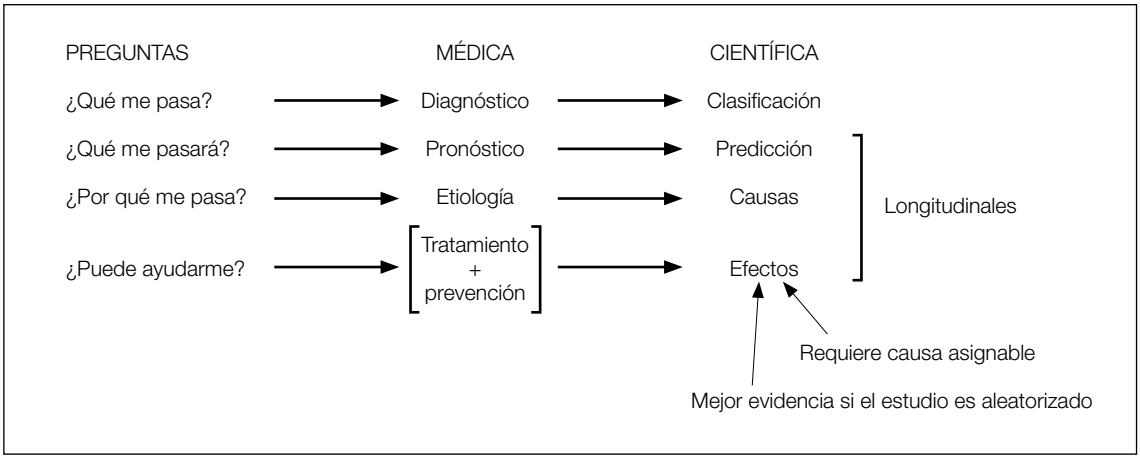


Figura 2. Preguntas clínicas y tipos de estudios.

El diagnóstico pretende una clasificación fina, en la cual los casos de un mismo grupo son similares entre sí pero diferentes de los de otros grupos. Los estudios de la precisión diagnóstica de un indicador lo enfrentan a una variable de referencia *(gold estandard)*. El pronóstico también pretende hacer grupos homogéneos, pero ahora respecto al futuro.

Preguntas de seguimiento frente a preguntas instantáneas

En un estudio diagnóstico, los datos sobre el indicador y sobre la referencia pueden recogerse simultáneamente, pero los estudios de predicción y los de intervención requieren un intervalo de tiempo. Cuando las dos variables en estudio se observan en el mismo momento se habla de estudios transversales; en cambio, cuando una acontece previamente a la otra, se habla de estudios longitudinales.

La relación causal también precisa un lapso de tiempo para que se manifieste el efecto. En la investigación etiológica puede buscarse la información en el pasado, pero la confirmación de efectos requiere asignar la causa y observar el efecto en el futuro. Por ejemplo, si cierto componente plasmático ha de predecir enfermedad cardiovascular, debe ser previo en el tiempo a ésta, ya que si la determinación analítica se realizara simultáneamente a la aparición de la enfermedad, el valor de anticipación sería nulo. Aun más, si el objetivo fuera especular si hipotéticos cambios en dicho componente modificarían la enfermedad cardiovascular, se requiere este periodo de tiempo para que se manifiesten dichos efectos.

Un estudio diagnóstico relaciona variables simultáneas, pero los pronósticos y los de intervención precisan un lapso de tiempo entre ellas.

Variables iniciales y finales

En un estudio predictivo, la variable inicial *(input)* será el índice o el indicador pronóstico. En un estudio de intervención, la variable inicial será la maniobra o el tratamiento que se aplica. En ambos tipos de estudios se observará, al final *(output, end-point)* del seguimiento, la respuesta *(outcome)* o desenlace.

Una vez iniciado un estudio, se desea conocer la evolución de todos los casos que cumplen los criterios de selección. Aunque es frecuente usar el término «criterios de exclusión» para hablar de pacientes que no deben ser incluidos, como puede ser malinterpretado en el sentido de poder excluir pacientes a lo largo del estudio, el punto 4ª de la declaración CONSORT aconseja hablar únicamente de criterios de elegibilidad. Evitar la pérdida de casos es tan importante que la revista *New England Journal of Medicine* ha recordado a sus autores que cualquier exclusión, pérdida o dato ausente aumenta la incertidumbre, y por tanto debería ser prevenida o tratada con un buen análisis.[3]

En los estudios de cohortes, el criterio que determina la elegibilidad de los pacientes es inicial. Si son un conjunto, se habla de criterios de inclusión o de selección. En resumen, si los casos de un estudio se seleccionan en función de una variable inicial, recibe el nombre de «estudio de cohortes». Si además sus casos se asignan al azar a varias opciones terapéuticas en comparación, se habla de «ensayo clínico».

Los estudios que validan un índice pronóstico o cuantifican los efectos de una intervención tienen una variable final que indica la evolución o el resultado. Por ser desconocida al inicio, es aleatoria en términos estadísticos. Pero en los estudios etiológicos, que buscan posibles causas, puede “invertirse” el orden de recogida de las variables. Así, la evolución (variable final o respuesta) es el criterio que determina la inclusión del individuo, y luego se investiga en el pasado el valor de las posibles causas que son ahora las variables en estudio (aleatorias en términos estadísticos). Por ejemplo, en un estudio de casos y controles se seleccionan unos casos con la enfermedad en estudio y unos controles que no la tienen, y se averigua su exposición previa a posibles causas hipotéticas.

En resumen, la variable que determina la inclusión del individuo puede ser inicial (cohortes y ensayos clínicos) o final (casos y controles). Por otro lado, la referencia o control en un ensayo clínico es una intervención (variable inicial), pero

en un estudio etiológico de casos y controles es una evolución (final). Obsérvese, por tanto, que el término «control» se aplica a la variable inicial o tratamiento en un ensayo clínico, y a la situación final en un estudio de casos y controles.

Hacer frente a ver

En los estudios experimentales, el investigador asigna el valor de la intervención a los voluntarios, pero en los estudios observacionales las unidades se presentan con valor en las variables de estudio. Por ejemplo, si se pretende estudiar el efecto de la monitorización de los pacientes hipertensos en el control de su presión, en un estudio observacional los médicos y los pacientes decidirán el número y el momento de las visitas, pero en un estudio experimental el investigador asigna un número de visitas a cada voluntario.

De este modo, la asignación permite distinguir entre experimentos y observaciones. Sin embargo, por respeto al principio de no maleficencia, sólo las intervenciones que pretendan mejorar el estado de salud son asignables. Por ejemplo, un adolescente no puede asignarse al grupo «fumador de tabaco desde los 15 hasta los 50 años». De aquí, la predilección de la epidemiología por la observación. En cambio, la pregunta habitual de la farmacología («¿mejora este tratamiento la evolución?») permite la asignación del tratamiento y, por tanto, el diseño experimental. Para recurrir a la asignación, primero la epidemiología debe redefinir la causa en estudio para convertir en positivos los efectos. Por ejemplo, ¿qué pasará si introduzco esta ayuda para dejar de fumar? La gran ventaja de la asignación es que permite utilizar las herramientas del diseño de experimentos para minimizar errores. Pero además también permite evaluar si, cuando se asigne la causa en estudio, las unidades seguirán el consejo. En el ejemplo anterior del seguimiento observacional de los pacientes hipertensos, la primera asunción necesaria para aplicar los resultados a una intervención futura es que los pacientes se visitarán con la frecuencia sugerida por el médico. En cambio, el estudio experimental permite observar y cuantificar hasta qué punto los destinatarios de la intervención siguen las recomendaciones.

No obstante, los estudios experimentales no ofrecen ventajas cuando se persiguen las otras finalidades de la investigación sanitaria. Por ejemplo, en la predicción, si se desea establecer el pronóstico, un estudio de seguimiento no experimental (longitudinal prospectivo o de cohortes) basado en un muestreo aleatorio representativo será mejor que un ensayo clínico con selectivos criterios de elegibilidad. Igualmente, si se desea valorar la capacidad diagnóstica de un indicador, el diseño instantáneo o transversal es suficiente. Hay que recordar, por tanto, que los mejores diseños para valorar las capacidades diagnóstica y pronóstica son los transversales y los de seguimiento, respectivamente.

Preguntas sobre efectos
y preguntas sobre causas

En el entorno de la relación causa-efecto conviene distinguir entre preguntas sobre efectos y preguntas sobre causas. Nótese la diferencia entre «si me tomo una *Aspirina,* ¿se me irá el dolor de cabeza?» y «se me ha ido el dolor de cabeza, ¿será porque me tomé una *Aspirina*?».

Al estudiar relaciones de causa-efecto también conviene distinguir si el objetivo es confirmar o explorar. El ensayo clínico aporta la evidencia de mayor calidad para confirmar y estimar un efecto, pero si el objetivo es explorar posibles causas, como en la investigación etiológica, muchas evidencias provendrán de estudios de cohortes o de casos y controles bien diseñados.

Así, el estudio de relaciones de causa-efecto suele comportar dos pasos sucesivos. El primero, dado un determinado efecto (una enfermedad, por ejemplo) se desea explorar sus posibles determinantes, sus causas. En el segundo paso, identificada una causa asignable, es decir, que puede ser intervenida, se desea confirmar y cuantificar el efecto que origina su intervención. Por ejemplo, cuando el Dr. Joan Clos encargó a los Dres. Jordi Sunyer y Josep Maria Antó estudiar las epidemias de asma en la Barcelona preolímpica, la respuesta a la primera pregunta, retrospectiva, «¿cuáles son las causas del

asma?», fue «la descarga de soja en el puerto con viento hacia el lugar de presentación de los casos». El estudio de aquello sobre lo que se podía intervenir y de aquello que, como el viento, no lo era, llevó a la segunda pregunta, prospectiva, «¿conseguiremos reducir los brotes de agudización del asma reparando el silo y protegiendo la descarga de soja?», que tuvo una respuesta positiva.

«Prospectivo» y «retrospectivo» son términos ambiguos

Un primer uso de prospectivo (P) y retrospectivo (R) hace referencia a la pregunta en estudio: sobre efectos (P) o sobre causas (R). Un segundo uso considera la estrategia de muestreo y recogida de datos, según si la variable que determina la inclusión en el estudio es inicial (P: cohortes, ensayo clínico) o bien final (R: casos y controles). Esta segunda acepción implica otra: que los datos sean futuros (P) o pasados (R); lo cual requiere recoger cada variable en el momento en que acaece (P) o bien buscando en el pasado la variable inicial (R). Finalmente, un tercer uso valora la existencia de una hipótesis independiente de los datos, llamando prospectivos a los estudios en que puede documentarse que la hipótesis es previa a la existencia de los datos (confirmatorios) y retrospectivos en caso contrario (exploratorios). Así, los términos «prospectivo» y «retrospectivo» tienen varios usos y acepciones, lo que quebranta uno de los principios fundamentales de la ciencia: «un término, un significado». Por tanto, conviene evitar, por su ambigüedad, los términos «prospectivo» y «retrospectivo». En su lugar, hay que aclarar si la pregunta es sobre causas o sobre efectos, si había hipótesis previa (confirmatorio o exploratorio) y cuál es la variable o criterio que determina la inclusión de un caso en el estudio.

Causas y condiciones

Intervenir implica cambiar algo, lo que requiere un mínimo de dos valores para la variable causa. Puede ser sustituir una opción terapéutica A por otra B, o añadir un nuevo tratamiento C a la guía clínica, o bien modificar los hábitos higiénico-dietéticos eliminando (o añadiendo) alguno.

Hay que insistir en la acción como intervención. Atributos como la edad o el sexo son útiles para hacer un pronóstico o una predicción; por ejemplo, cabe esperar que una mujer viva alrededor de 5 años más que un hombre. Pero no son modificables y, por tanto, no tiene sentido actuar (intervenir) sobre ellos. No tendría ningún sentido que un médico dijera «cambie usted este mal hábito de ser hombre; hágase mujer y vivirá 5 años más».

En consecuencia, desde un punto de vista práctico, de intervención, es irrelevante preguntarse si el sexo o la edad tienen un efecto causal en, por ejemplo, la supervivencia. Basta con conocer su capacidad pronóstica para anticipar el futuro. Ahora bien, el siguiente estudio “de laboratorio” permite estimar el efecto del sexo en el salario: se pregunta a empleadores por el sueldo que darían a los currículos de una serie de trabajadores a quienes, artificialmente, se les asigna el sexo al azar.

Principios estadísticos

Niveles de evidencia

Suele usarse una gradación sobre la calidad de evidencia que cada tipo de estudio puede aportar sobre una intervención: ensayos clínicos, estudios observacionales longitudinales, estudios observacionales transversales, y notificaciones anecdóticas de casos. Sin embargo, debe quedar claro que esta gradación de la evidencia se aplica a la intervención, pero no al diagnóstico ni al pronóstico. Por último, disponemos del metaanálisis, que es la técnica estadística que permite agregar la información contenida en los estudios. Una revisión sistemática formal con su correspondiente metaanálisis aportará una visión más global que la de estudios separados.

Determinismo frente a variabilidad

Para afirmar que mañana se hará de día sólo necesitamos asumir igualdad entre pasado y futuro. Pero si nos preguntamos si lloverá, además de asumir igualdad entre pasado y futuro necesita-

mos modelizar de qué depende la lluvia, tratar la variabilidad y cuantificar la duda. En resumen, considerar la variabilidad implícita en un proceso obliga a recurrir a la estadística, pero si el proceso carece de variabilidad puede olvidar la estadística.

Objetivos e hipótesis

Un objetivo es la motivación o finalidad subjetiva del estudio. La hipótesis, en cambio, expresa sin ambigüedades y de forma cuantitativa el criterio o consecuencia contrastable en que se basará la conclusión. Por ejemplo, «nuestra intención, según figura en el protocolo, es demostrar que añadir el componente T a la guía de práctica médica mejora la evolución de los pacientes. Nuestra hipótesis es que el nivel de autonomía, valorado a las 12 semanas con la escala de Rankin por evaluadores acreditados, tiene un valor mejor en los tratados que en los controles».

Hipótesis y premisas

No todas las ideas especificadas en el modelo conceptual del estudio tienen la misma importancia. El nombre «premisas» suele reservarse para las ideas acompañantes que son necesarias para poder contrastar las hipótesis. Por ejemplo, para estudiar el efecto de un nuevo tratamiento es usual asumir que el efecto es el mismo (constante) en todos los pacientes y que la evolución es independiente de un paciente a otro. La primera premisa podría verse afectada en un ensayo clínico con criterios de elegibilidad excesivamente amplios, y la segunda en una intervención grupal, como un consejo profiláctico en una clase de adolescentes o los efectos de una vacuna, cuando la probabilidad de contagio puede depender del efecto en otros casos.

Así, el objetivo principal de un estudio confirmatorio es contrastar con datos empíricos la hipótesis, aceptando o asumiendo como razonables ciertas premisas. Eso sí, analizar el grado de verosimilitud de las premisas constituye uno de los objetivos secundarios. En resumen, más relevante que saber si las premisas son ciertas es conocer que se llega a la misma conclusión si se parte de otras premisas.

Multiplicidad

Para controlar la posibilidad de obtener resultados simplemente por azar, el proceso habitual consiste en definir una sola hipótesis que se contrastará en una variable respuesta con un único método de análisis. La existencia de un protocolo público, escrito antes de acceder a los resultados, garantiza que se ha respetado el orden requerido en los estudios confirmatorios: primero la hipótesis y el plan estadístico, luego los datos, y finalmente el análisis.

Error aleatorio frente a error sistemático

Las clases de estadística empiezan con la frase «sea X una variable aleatoria de la que tenemos una muestra aleatoria». A partir de aquí se derivan métodos para cuantificar la posible influencia del azar y cuantificar la incertidumbre o el ruido del muestreo que, aplicado a la señal obtenida, proporciona estimaciones por intervalo de los valores poblacionales. Por ejemplo, en una muestra aleatoria de 2000 afiliados a un proveedor de servicios sanitarios se ha observado una proporción de un 20% de hipertensos (400/2000). Con una confianza del 95%, la auténtica proporción poblacional es algún valor comprendido entre 18,3% y 21,8%.

Por tanto, la estadística proporciona instrumentos para cuantificar la incertidumbre originada por un proceso aleatorio. Sin embargo, si la muestra no es aleatoria hay que recordar que existen otras fuentes de error no contempladas por las herramientas estadísticas. Por ejemplo, si se observa que un 50% (50/100) de los casos de botulismo registrados en cierta comunidad en cierto periodo fallecieron, ¿cómo cuantificar la incertidumbre? Es necesario considerar sus dos fuentes, aleatoria y no aleatoria, en dos pasos sucesivos. Para el primero, se asume que todos los habitantes de esa comunidad tienen la misma probabilidad de contraer botulismo. Si además se asume que dichas probabilidades son independientes entre sí, ya se dispone de los mecanismos que hubieran originado una muestra aleatoria simple y puede magnificarse el error aleatorio mediante un intervalo de confianza: cierto cálculo adecuado para muestras pequeñas (basado en

la distribución binomial) dice que, si los 100 casos proceden al azar de una población, observar 50 muertes es compatible con probabilidades de morir en la población comprendidas entre el 39,83% y el 60,17%, con una confianza del 95%.

El segundo paso es cuestionarse si se detectaron todos los casos de botulismo. Si, por ejemplo, cabe esperar que la mitad de las muertes por botulismo no fueran diagnosticadas como tales, deberíamos añadir 50 casos al numerador y al denominador, subiendo la mortalidad al 66% (100/150). En cambio, si lo que cabía esperar es que los casos leves no se diagnosticaran y su número se estima igual al de los casos diagnosticados, ahora deben añadirse 50 casos, pero sólo al denominador, resultando en una mortalidad del 33% (50/150). Este ejemplo muestra que la magnitud del error sistemático por imprecisiones en la recogida de los datos (del 33,33% al 66,67%) puede ser mayor que el error contemplado por un proceso aleatorio puro (intervalo de confianza del 95%: 39,83% a 60,17%).

El error originado por una obtención no aleatoria de los datos puede ir en cualquier sentido, por lo que se denomina «sesgo impredecible». De hecho, una monografía de Jon J. Deeks[4] para la agencia de tecnología sanitaria inglesa muestra que los estudios no aleatorizados tienen una mayor imprecisión que no contemplan las medidas estadísticas de error aleatorio ni se corrige con las técnicas de ajuste.

Saber (ciencia) frente a hacer (técnica)

Para interpretar correctamente los resultados hay que distinguir entre el objetivo científico de adquisición de conocimiento y el objetivo clínico que requiere tomar decisiones. De hecho, aumentar el conocimiento disponible requiere inducción, pero aplicarlo es un ejercicio deductivo: la inferencia adquiere conocimiento valorando las pruebas científicas (evidencia) a favor o en contra de los modelos establecidos. Para ello se recomienda utilizar los intervalos de confianza, aunque si el autor así lo considera puede reportar el valor p. Por otro lado, el acto médico, las medidas de salud pública, la gestión de recursos o el permiso para comercializar un nuevo fármaco implican un proceso de decisión con riesgos asociados a dos posibles acciones erróneas contrapuestas (errores de tipo I y tipo II).

Como el conocimiento en sí mismo no tiene implicaciones, pero sí las acciones y las decisiones que se toman basándose en él, debe distinguirse entre “almacenes” de conocimiento (revistas científicas, bibliotecas o las colaboraciones Cochrane y Campbell) y órganos de decisión (agencias reguladoras, departamentos de farmacia, agencias de salud pública). El proceso de decisión incluye la inferencia, pero también las opiniones sobre los posibles resultados: utilidad, coste, preferencias o cualquier función de pérdida. Un ejemplo muy celebrado es que antes de usar el paracaídas en un salto desde mil metros de altura nadie preguntaría por el ensayo aleatorizado y enmascarado que aporte las pruebas científicas sobre el efecto beneficioso del paracaídas.

Por supuesto, las decisiones implícitas en el acto médico deben basarse en el conocimiento contrastado empíricamente que aportan los artículos científicos, pero también en las consecuencias (utilidades, beneficios, costes, etcétera) de las alternativas en consideración, y su valoración por los destinatarios. Como todas estas últimas pueden variar fácilmente de un entorno a otro, es más fácil establecer un conocimiento común que recomendar acciones generales para todo el universo. Precisamente, la teoría de la decisión aporta los elementos para racionalizar el paso desde un artículo científico “universal” a una guía de práctica clínica “local”.

Bibliografía

1. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. Disponible en: http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124
2. Jager LR, Leek JT. Empirical estimates suggest most published medical research is true. Disponible en: http://arxiv.org/ftp/arxiv/papers/1301/1301.3718.pdf
3. Ware JH, Harrington D, Hunter DJ, D’Agostino RB. Missing data. N Engl J Med. 2012;367:1353-4.
4. Deeks JJ, Dinnes J, D’Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technology Assessment. 2003;7(27). Disponible en: http://www.hta.ac.uk/fullmono/mon727.pdf

Hay más casualidades que causalidades

González / Ventura

La epidemiología y los estudios observacionales de cohortes y de casos y controles

José Luis Peñalvo

La epidemiología es el estudio de la aparición de enfermedades en poblaciones humanas a partir del recuento de los episodios relacionados con la salud que presentan las personas en la relación con los grupos que existen de forma natural (poblaciones) y a los cuales pertenecen dichas personas.[1] La epidemiología clínica es una de las ciencias básicas en que se apoyan los médicos para la asistencia a los pacientes, y tiene como objetivo el desarrollo y la aplicación de métodos de observación clínica que den lugar a conclusiones válidas, evitando las equivocaciones derivadas del error sistemático y del azar.[2] La medicina basada en la evidencia es un concepto relativamente reciente, que se refiere a la aplicación de la epidemiología clínica a la asistencia de los pacientes y que incluye, entre otros aspectos, la revisión crítica de la información derivada de las investigaciones epidemiológicas para la toma de decisiones.[3]

El método epidemiológico

El método epidemiológico puede dividirse en fases. La primera de ellas, la fase descriptiva, tiene como objetivo la caracterización de los problemas de salud que existen en una determinada población, qué impacto tiene un problema determinado en dicha población y cuáles son sus patrones de aparición en cuanto a tiempo, lugar y población afectada. La epidemiología descriptiva nos sirve para componer una fotografía fija de un problema de salud pública. Dentro de esta área descriptiva, los estudios de prevalencia examinan a las personas que forman parte de una población en busca de un efecto o suceso de interés en un tiempo determinado. La fracción de población que presenta el efecto constituye la prevalencia de la enfermedad. Este tipo de estudios se denominan también estudios transversales o estudios de corte, ya que las perso-

nas son estudiadas en un momento del tiempo. Nos proporcionan información sobre la carga de la enfermedad y la magnitud de ésta, es decir, en una analogía con el periodismo, estos estudios contestan a las preguntas de quién, dónde y cuándo. Posteriormente, al definir la exposición (posible causa de la enfermedad o suceso) las preguntas serán por qué y cómo, y al definir el efecto la pregunta será qué. Hay tres indicadores básicos a la hora de estudiar la carga de una enfermedad: 1) la prevalencia, definida como el número de casos de una enfermedad determinada o exposición en una población y en un momento dado; 2) la incidencia, o los casos nuevos de una enfermedad en una población definida dentro de un plazo determinado (también denominada densidad de incidencia o tasa de incidencia), y 3) la mortalidad, o número de defunciones en una población por cada 1000 habitantes durante un periodo determinado (generalmente 1 año).

Una vez que el problema de salud pública se ha caracterizado, es función de la epidemiología generar hipótesis que expliquen los patrones hallados. La generación de hipótesis y su evaluación no dependen sólo de las observaciones realizadas, sino también del conocimiento de los resultados de estudios previos, así como de la integración del conocimiento de otras áreas científicas y, por último, de la intuición del investigador. El epidemiólogo pretende contestar a las mismas preguntas que un periodista, y el resultado final de un trabajo epidemiológico se podría comparar a generar una noticia. En el estudio de hipótesis se busca, generalmente, la causalidad. Al método para determinar las razones y causas de los datos observados se le denomina «inferencia causal», y es uno de los principales objetivos de la epidemiología. En esta fase de inferencia, se denomina «variable» (aquello que varía y puede medirse) a los atributos de los pacientes y a los episodios clínicos, y en un estudio típico de asignación de causalidad hay tres tipos de variables: una es una supuesta exposición o variable predictora (a veces denominada variable independiente), la otra es un efecto o evento (a veces denominada variable dependiente) y la tercera es un grupo de variables que pueden afectar a la relación de las dos primeras, y se denominan variables extrañas, covariables y, en algunos casos, variables confusoras.

En resumen, el método epidemiológico es un proceso de estudio de hipótesis que expliquen los patrones de distribución de las exposiciones y los sucesos observados, eliminando aquellas que no sean concordantes con las observaciones. Así, las fases del proceso son: 1) generación de hipótesis; 2) diseño e implementación de estudios para generar variables y observaciones; 3) descripción de la distribución de las observaciones (análisis exploratorio de datos), y 4) inferencia o evaluación de la magnitud de la evidencia (análisis de datos confirmatorio).

Estudios de cohortes

La población analizada en un estudio de incidencia se denomina «cohorte». Una cohorte se define como un grupo de personas que tienen una característica en común en el momento en que se forma el grupo, y que son seguidas prospectivamente en el tiempo hasta que se produce el suceso objeto de estudio. Los estudios de cohortes también se conocen como longitudinales (participantes seguidos a lo largo del tiempo) y prospectivos (dirección del seguimiento, hacia delante). En este diseño, el riesgo o la incidencia de la enfermedad y el riesgo relativo (incidencia en expuestos frente a incidencia en no expuestos) se obtienen directamente.

En los estudios de cohortes se selecciona a los participantes en función de que presenten o no la exposición que interesa al investigador, quien a posteriori buscará el efecto propuesto en su hipótesis. Una cohorte es, por tanto, un grupo de personas definidas en un espacio de tiempo y un lugar determinado, con unas exposiciones definidas y en quienes los epidemiólogos tienen un efecto claro que buscar, asociado a esas exposiciones. La principal medida que se utiliza en un estudio de cohortes es el riesgo relativo: el número de veces que la incidencia de un efecto es mayor en los sujetos expuestos que en los no expuestos.

Las principales fases de un estudio de cohortes son: 1) selección de los participantes; 2) obtención de datos de la exposición; 3) se-

guimiento de los participantes, y 4) análisis de los datos.

Este tipo de estudio tiene algunas ventajas respecto a los experimentales. En primer lugar, permite estudiar la progresión de una enfermedad en el caso de que, por distintos motivos (entre ellos, éticos), no puedan llevarse a cabo estudios experimentales. Esto permite ver los diferentes factores que, a lo largo del estudio, van influyendo en el desarrollo o la evolución de la patología. Este tipo de estudio observacional permite, frente a los experimentales, demostrar varias hipótesis a la vez.

Los estudios de cohortes son, seguramente, la manera más lógica de estudiar la aparición de sucesos, aunque en la práctica son complicados de llevar a cabo y requieren una gran inversión. Las enfermedades crónicas tardan mucho tiempo en manifestarse, y el periodo de latencia desde la exposición hasta el desarrollo de síntomas es largo. Por tanto, la principal ventaja de un estudio de cohortes es su duración. Esto supone un alto remplazo del personal técnico e investigador del estudio, por lo que se necesita un altísimo grado de estandarización de los procedimientos y métodos para que la recogida de datos y su análisis sean siempre homogéneos y los datos sean concordantes durante el tiempo que dure el estudio. En una cohorte, además, también es fácil que se introduzca un sesgo de selección a la hora de escoger a los participantes. Por último, sin ser una desventaja, hay que tener en cuenta que el análisis estadístico de este tipo de estudios es muy complejo y requiere un alto grado de conocimientos.

Un ejemplo clásico de este tipo de estudios es el Framingham, iniciado en la década de 1940 y del que se ha obtenido la caracterización de los principales determinantes del riesgo de aterosclerosis.[4] En el caso del estudio Framingham las exposiciones eran los factores de riesgo clásicos de enfermedad cardiovascular (hipertensión, colesterol…), y la hipótesis, demostrada, es que dichos factores de riesgo aumentan las posibilidades de padecer aterosclerosis.

En la actualidad, el Centro Nacional de Investigaciones Cardiovasculares coordina la cohorte del PESA *(Progression of Early Subclinical Atherosclerosis)*.[5] Este estudio de cohortes ayudará a identificar los factores de riesgo que influyen en el desarrollo de la aterosclerosis, y mejorará así la prevención de la enfermedad aterosclerótica al poder hacer un diagnóstico precoz incluso antes de la aparición de los síntomas. Hasta ahora, diversos proyectos han intentado evaluar si las técnicas avanzadas de diagnóstico por la imagen pueden ayudar a la detección precoz de la enfermedad, pero la mayoría se llevan a cabo en poblaciones mayores de 60 años. Se ha demostrado que, en este grupo de edad, la aterosclerosis lleva ya décadas desarrollándose, por lo que podría no ser reversible aunque se detectara de forma precoz. Este estudio usa las técnicas de diagnóstico por la imagen más avanzadas de Europa para identificar individuos con aterosclerosis subclínica, en una población de 4500 trabajadores de 40 a 54 años de edad. El estudio examinará la asociación entre los parámetros clínicos y la presencia de factores genéticos, epigenéticos, metabolómicos, proteómicos y ambientales, incluyendo los hábitos dietéticos, la actividad física, las características psicosociales y la exposición a contaminantes ambientales.

Estudios de casos y controles

Una alternativa a los diseños de alto coste económico como los de cohortes, muy utilizada, son los denominados estudios de casos y controles. En ellos no es necesario esperar a que tras la medición de la exposición se produzca el suceso, sino que éstos (los casos) son seleccionados por el investigador entre un grupo de pacientes disponibles y, paralelamente, se selecciona un grupo de individuos sanos que servirán como controles. Una vez identificados los casos y los controles, la exposición se mide o se identifica de manera retrospectiva.

En este diseño no hay forma de conocer las tasas del suceso, ya que los grupos no se determinan de forma natural (como en una cohorte) sino en función de los criterios de selección del investigador. Por tanto, no puede calcularse una tasa de incidencia entre el grupo de personas expuestas y no expuestas, ni determinar el riesgo relativo dividiendo la incidencia entre personas expuestas y no expuestas. Sin embargo, lo que

sí puede calcularse es la frecuencia relativa de exposición en los casos y los controles, que proporciona una medida de riesgo conceptualmente similar al riesgo relativo. Es lo que se conoce como razón de posibilidades *(odds ratio),* que se define como la posibilidad de que un caso esté expuesto dividida por la posibilidad de que un control lo esté.

Si la frecuencia de exposición es más alta entre los casos, la *odds ratio* será superior a 1, lo que indicará un mayor riesgo. Por tanto, cuanto más estrecha sea la asociación entre exposición y enfermedad, mayor será la *odds ratio,* y viceversa. La interpretación es, de esta forma, similar al riesgo relativo obtenido de estudios de cohortes, y de hecho ambas medidas son equivalentes cuando la incidencia de la enfermedad es reducida.

Asignación de causalidad: estudios controlados y aleatorizados frente a estudios observacionales

En la evaluación de la eficacia de un tratamiento o intervención se busca estudiar la magnitud de la asociación entre una exposición y un efecto. Podría decirse que el mayor grado de asociación es la relación causal, en la cual la exposición es la causa del efecto observado. En epidemiología es comúnmente aceptado que el diseño más apropiado para establecer una relación causal es el del ensayo controlado (en el cual la exposición bajo evaluación se compara frente a un control, normalmente placebo), con asignación aleatoria de los pacientes a uno u otro grupo. El objetivo de esta aleatorización es evitar la influencia de las diferentes características individuales en los resultados finales, promoviendo que éstas se repartan de forma similar (aleatoria) entre los grupos a comparar. El objetivo final es, por tanto, que los grupos sólo difieran en la característica objeto de análisis.

En el caso de los denominados estudios observacionales, los participantes no son asignados a ningún tratamiento o, si lo son, la asignación se efectúa según la práctica clínica habitual y el investigador es un simple observador de la aparición o del desarrollo del efecto que es objeto de estudio.

Aunque en principio un diseño de este tipo se considera el estándar para la asignación de causalidad, también se hacen objeciones a la idoneidad de un ensayo clínico aleatorizado para asignar causalidad, ya que muchas veces este tipo de estudios conllevan criterios de inclusión muy restrictivos y poblaciones definidas (hospitales universitarios, centros de referencia, etcétera) que ponen en duda la generalización de los resultados obtenidos. Por tanto, lo más adecuado es pensar en ambas opciones como diseños epidemiológicos complementarios, ya que los estudios observacionales pueden servir para verificar y replicar resultados procedentes de estudios controlados y aleatorizados en la práctica habitual, en los que la variabilidad de las características individuales es mayor.

Los diseños observacionales también son los más adecuados para el estudio de efectos o sucesos muy poco frecuentes, o de tratamientos largos, en los que el reclutamiento de un grupo control no es éticamente aceptable, y en general para todas aquellas hipótesis en que la experimentación no es posible y sólo podemos inferir relaciones causales a partir de la observación.

Si bien es cierto que hay diferencias relevantes entre los ensayos aleatorizados y los estudios observacionales en cuanto a la magnitud de la relación (causalidad) entre exposición y efecto, la elección de un diseño u otro depende de las características de la hipótesis de partida a estudiar, y elegir un diseño controlado y aleatorizado no es sinónimo de garantía de calidad. La asignación de causalidad mediante estudios experimentales u observacionales es un tema que ha generado un amplio debate durante décadas. Actualmente, uno de los criterios más mencionados como estándar de definición de causalidad son los rasgos teóricos para valorar una asociación como signo de causalidad propuestos por el conocido epidemiólogo Austin Bradford Hill,[6] que son, de forma resumida:

- Temporalidad: la causa precede al efecto.
- Solidez: riesgo relativo grande.
- Dosis-respuesta: a mayor exposición, mayor tasa de enfermedad.

- Reversibilidad: la reducción en la exposición disminuye la tasa de enfermedad.
- Consistencia: observaciones homogéneas en distintas circunstancias, lugares, momentos y personas.
- Plausibilidad: sentido biológico.
- Especificidad: una exposición conlleva un efecto.
- Analogía: la causalidad ya está establecida para una relación exposición-efecto similar.

Quizá los más relevantes para un estudio observacional son la solidez (fuerza de la asociación), que viene determinada por el diseño epidemiológico y el modelo estadístico utilizado, y la evidencia experimental. En epidemiología observacional es muy importante recoger covariables que describan exposiciones coincidentes con la exposición objeto de estudio, así como la máxima información posible para luego poder ajustar la asociación que se está estudiando hasta prácticamente poder decir que la exposición tiene cierto grado de causalidad, casi como en un estudio experimental.

Confusión y sesgos

El sesgo es un proceso que, durante la etapa de inferencia, introduce de forma sistemática desviaciones en la recogida, el análisis, la interpretación, la publicación o la revisión de los datos de un estudio, y puede dar lugar a conclusiones que difieren de la verdad.[7] Hay tres tipos generales de sesgos: 1) sesgo de selección, derivado de la comparación de grupos de individuos que difieren en factores determinantes para el resultado, pero que no han sido objetivo del diseño; 2) sesgo de medición, originado cuando los instrumentos de medición difieren entre los participantes, y 3) sesgo de confusión, que se produce cuando los factores están asociados y el efecto de uno influye en el efecto del otro.

Los sesgos de selección y confusión no son mutuamente excluyentes, pero se describen por separado porque representan diferentes problemas en el transcurso de un estudio epidemiológico. El sesgo de confusión debe tenerse en cuenta durante el análisis de los datos, mientras que el sesgo de selección es crucial durante la fase de diseño del estudio.

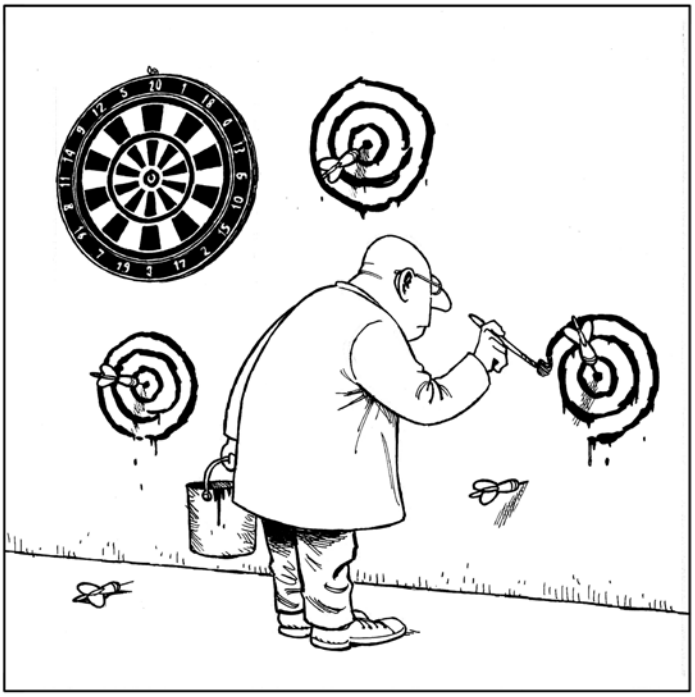
La principal diferencia entre los ensayos controlados aleatorizados y los estudios observacionales radica, por tanto, en la variabilidad interindividual en la población objeto de estudio. En los primeros, el objetivo del diseño permite disminuir al mínimo esta variabilidad para asignar causalidad al tratamiento; en los segundos, el reclutamiento de los participantes normalmente está sujeto a posibles sesgos de selección. Por ejemplo, cuando los participantes incluidos en una cohorte difieren en otras características además de la variable de estudio (presencia de otras enfermedades, por ejemplo), se produce un sesgo de selección denominado «de susceptibilidad». Puede producirse un sesgo de migración cuando los integrantes de un subgrupo de la cohorte abandonan este grupo para pasar a otro (por ejemplo, trabajadores activos pasan a inactivos con el transcurso de los años).

Existen métodos para corregir, en cierta medida, los sesgos originados durante el diseño, el desarrollo y el análisis de un estudio epidemiológico:

- Aleatorización: pacientes elegidos al azar para formar parte de un grupo determinado.
- Restricción: limitación en la variabilidad de las características de la población.
- Emparejamiento: a cada participante de un grupo se le asigna uno (o más) participantes con las mismas características, excepto por la variable objeto de estudio, con el fin de crear comparaciones.
- Estratificación: comparación de resultados entre subgrupos que tienen probabilidades similares de obtener un mismo resultado.
- Ajuste sencillo: ajusta matemáticamente las tasas brutas o crudas en función de variables seleccionadas para dar el mismo peso a subgrupos con riesgos similares.
- Ajuste multivariable: ajusta, mediante modelos matemáticos, las diferencias de diferentes factores asociados con el resultado.

Bibliografía

1. Friedman GD. Primer of epidemiology. 5th ed. New York: Appleton and Lang; 2003.
2. Fletcher RH, Fletcher SW. Epidemiología clínica. 4ª ed. Barcelona: Wolters Kluwer; 2008.
3. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine. How to practice and teach EBM. 2nd ed. New York: Churchill Livingstone; 2000.
4. Dawber TR. The Framingham Study: the epidemiology of aterosclerosis disease. Cambridge, MA: Harvard University Press; 1980.
5. Estudio PESA (Progression of Early Subclinical Atherosclerosis). Disponible en: www.estudiopesa.org
6. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965;58: 295-300.
7. Last JM. A dictionary of epidemiology. 3rd ed. New York: Oxford University Press; 1995.



Las hipótesis deben ser previas a los resultados del estudio

Schwartz-Woloshin / Ventura

La confianza en los resultados de la investigación y el sistema GRADE

Pablo Alonso Coello

Introducción

Para tomar decisiones necesitamos estar adecuadamente informados. En las que afectan a nuestra salud, o simplemente a la hora de poder valorar una información proveniente de la investigación, necesitamos saber hasta qué punto podemos confiar en los resultados disponibles. Esta confianza también se ha denominado habitualmente calidad o riesgo de sesgo. Más recientemente, el concepto de calidad (o confianza) ha evolucionado, y en él se incluyen ahora otros factores, que van desde el diseño y la ejecución de los estudios hasta la precisión de los resultados, entre otros *(quality of evidence).* Como veremos, la calidad es, por tanto, un atributo continuo, y nuestro veredicto dependerá de la ponderación de estos factores.

¿Cómo podemos evaluar la calidad?

Los profesionales de la salud, a través de sus sociedades científicas principalmente, y las instituciones sanitarias en general, vienen evaluando la calidad de la información mediante diferentes sistemas desde hace ya más de dos décadas.[1] Estos sistemas que evalúan la calidad –y algunos de ellos también la fuerza de las recomendaciones– se han denominado sistemas de clasificación de la evidencia o, más comúnmente, niveles de evidencia. Las recomendaciones son el ingrediente fundamental de las guías de práctica clínica, un instrumento cada vez más popular para la toma de decisiones de los profesionales sanitarios.[2] Para realizar recomendaciones se tiene en cuenta no sólo la calidad de la información, sino también otros factores como, por ejemplo, el balance beneficio-riesgo, las preferencias o los costes.[3]

Un aspecto que ha generado confusión es la convivencia de diferentes sistemas que presentan, en mayor o menor medida, similares limitaciones. En general, una de las más frecuentes es la penalización de los estudios observacionales, asignándoles evidencia de calidad baja.[1,4] Esto

es contraintuitivo, pues hay bastantes situaciones en las que no es necesario realizar un ensayo clínico para tener una confianza alta en el efecto de una intervención. Por ejemplo, cuando se descubrió la insulina, el efecto de su administración en unos pocos casos fue tan espectacular que no se llegaron a realizar ensayos clínicos, ni ya nunca se realizarán. Por otra parte, a pesar de disponer de ensayos clínicos para una intervención, la confianza puede ser baja debido a factores como, por ejemplo, las limitaciones en su diseño y ejecución.

¿Dónde encajan las revisiones sistemáticas?

Otra de las limitaciones de estos sistemas de clasificación es que otorgaban un peso excesivo a las revisiones sistemáticas. Este tipo de revisiones son el método de referencia a la hora de conocer los efectos de las intervenciones, pero debemos tener en cuenta que sus resultados son tan fiables como los estudios que incluyen.[5,6] Por ello, nuestra confianza debe estar fundamentada en el conjunto de los estudios que contiene una revisión sistemática (contenido) y en su calidad, y no en el hecho de que sea una revisión sistemática (continente).

Archie Cochrane señaló en 1972 la necesidad de realizar un mayor número de revisiones, cuando afirmó que era muy grave que todavía no se hubiese organizado una síntesis crítica de todos los ensayos clínicos relevantes, por especialidades o subespecialidades, y que fuera actualizada con periodicidad.[7] Precisamente, la respuesta a este desafío es la Colaboración Cochrane, una organización internacional sin ánimo de lucro que tiene como objetivo principal ayudar a tomar decisiones clínicas y sanitarias bien fundamentadas, preparando, manteniendo y divulgando revisiones sistemáticas sobre los efectos de la atención sanitaria (http://www.cochrane.org/).[8,9]

Las revisiones sistemáticas son investigaciones (investigación secundaria) que sintetizan los resultados de un conjunto de estudios individuales (investigación primaria). Las características fundamentales que mejor las definen son: 1) realizan una síntesis y un análisis de la información; 2) están basadas en la mejor evidencia científica disponible; 3) formulan preguntas claramente definidas, y 4) utilizan métodos sistemáticos y explícitos para identificar y seleccionar estudios, evaluarlos críticamente, extraer los datos de interés y analizarlos.[6] En consecuencia, las revisiones sistemáticas tienen como objetivo ser: 1) rigurosas, evaluando la calidad de los estudios incluidos; 2) informativas, esto es, enfocadas hacia problemas reales, presentando la información del modo que mejor ayude a la toma de decisiones; 3) exhaustivas, con el objetivo de identificar y utilizar la mayor cantidad posible de información pertinente, minimizando la introducción de posibles sesgos (por ejemplo sesgo de publicación), y 4) explícitas, con los métodos utilizados descritos de manera detallada.[6]

Revisiones sistemáticas frente a metaanálisis

En ocasiones, una revisión sistemática puede incluir un metaanálisis, un concepto introducido por Glass en 1976 y que se define como una integración estructurada, con una revisión cualitativa y cuantitativa, de los resultados de diversos estudios independientes acerca de un mismo tema.[5] En otras palabras, un metaanálisis es la combinación estadística de al menos dos estudios para obtener una estimación global sobre el efecto de una intervención. En la figura 1 podemos ver la representación de un metaanálisis estándar. Los estudios individuales están indicados con un cuadrado (efecto) y una línea horizontal (intervalo de confianza), de tal modo que cuanto más corta es la línea más preciso es el resultado. Finalmente, abajo, con un rombo, se representa el efecto global de la intervención resultante del metaanálisis. La línea vertical que pasa por el número 1 indica la posición alrededor de la cual los resultados se concentrarían si las dos intervenciones comparadas tuviesen efectos similares. Si una línea horizontal toca esta línea, significa que aquel ensayo clínico concreto no halló diferencias claras entre los tratamientos. La posición del rombo a la izquierda de esta línea indica que el tratamiento estudiado es beneficioso, y a la derecha perjudicial.

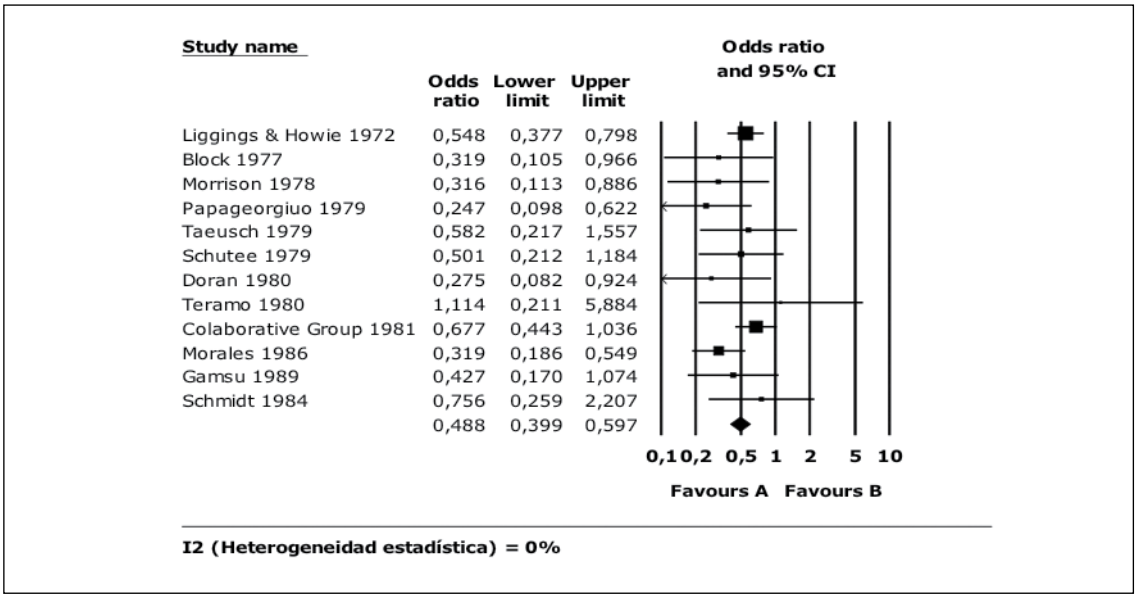


Figura 1. Efecto de los corticosteroides antenatales sobre el distrés respiratorio neonatal.[10]

El logotipo de la Colaboración Cochrane (figura 2) ilustra el metaanálisis de la figura 1, que corresponde a una revisión sistemática que evaluó la administración de corticosteroides (un tratamiento corto y barato) a mujeres gestantes con amenaza de parto prematuro, en comparación con placebo, para prevenir la disnea.[10] En la revisión se incluyeron 12 ensayos clínicos que evaluaban esta misma cuestión, el primero de ellos publicado en 1972. La revisión proporcionó una síntesis de la información disponible una década después de la publicación del primer ensayo que mostraba que los corticosteroides reducen el riesgo de muerte de los recién nacidos. En 1991 ya se habían realizado siete ensayos más, que mostraron aún con más claridad que esta intervención reduce un 30% a un 50% la probabilidad de morir de los recién nacidos.

Figura 2. Logotipo de la Colaboración Cochrane.

Sin embargo, como no se publicó ninguna revisión sistemática de estos ensayos hasta 1989, la mayoría de los obstetras no pudo conocer la efectividad real de este tratamiento, y decenas de miles de recién nacidos prematuros no se beneficiaron de recibir esta intervención. Éste es uno de los muchos ejemplos que ponen de manifiesto la necesidad de disponer de revisiones sistemáticas actualizadas sobre todos los aspectos de la atención sanitaria.

Aunque a veces se utilicen indistintamente los dos términos, revisión sistemática no es equivalente a metaanálisis. Podemos tener una revisión sistemática sin metaanálisis, pero no deberíamos tener un metaanálisis sin revisión sistemática. En el primer caso, los resultados de los estudios primarios se resumen, pero no se combinan con métodos estadísticos (revisión sistemática «cualitativa»).[5,6] Si se realiza un metaanálisis, estamos ante una revisión sistemática «cuantitativa». El metaanálisis es, por tanto, sólo una parte, aunque importante, de la revisión sistemática. La virtud fundamental del metaanálisis es que aumenta la potencia estadística del análisis, pues

disponemos de más estudios y pacientes, y por tanto proporciona resultados más precisos.

Las revisiones cumplen otra función fundamental: dan a conocer de un solo vistazo toda la información disponible sobre una pregunta de investigación. Esto es crucial desde el punto de vista del periodismo científico, pues ayudan a contextualizar los resultados de los nuevos estudios. Además, sirven para señalar las deficiencias de la investigación realizada hasta el momento y la investigación que es necesaria en un determinado campo. Por tanto, bienvenidas sean también para los periodistas. Así pues, conviene extremar las cautelas cuando en un comunicado de prensa o en un artículo no se contextualizan los resultados con la investigación previa, idealmente mediante una revisión sistemática. No obstante, muchos ensayos todavía no empiezan ni terminan con una revisión sistemática, como se recomienda.[11]

¿Cómo valorar una revisión sistemática?

Para leer una revisión sistemática tenemos que comprobar al menos dos factores clave. El primero es que presente una pregunta y unos criterios de inclusión y exclusión de los estudios claros y concretos. El segundo, que documente haber realizado una búsqueda de la literatura disponible, idealmente en una o más bases de datos o fuentes. Y el tercero, que haya realizado una evaluación de al menos el diseño y la ejecución de los estudios, esto es, que haya evaluado de manera crítica el contenido. Existen instrumentos más sofisticados para evaluar la calidad de las revisiones sistemáticas, pero llevan tiempo y requieren cierta experiencia para ser aplicados. En el contexto del periodismo, en principio es suficiente poder reconocer una revisión sistemática y diferenciarla de la que probablemente no lo es.

El sistema GRADE

En este contexto, un grupo internacional de epidemiólogos, metodólogos y clínicos ha desarrollado una propuesta que tiene como objetivo consensuar un sistema común para evaluar la calidad o confianza, y que supere las limitaciones de los sistemas previos. Este grupo de profesionales ha constituido el grupo de trabajo GRADE *(Grading of Recommendations Assessment, Development and Evaluation).*[3,12] El sistema GRADE ha sido adoptado por más de 70 organizaciones en todo el mundo, algunas tan importantes como la Organización Mundial de la Salud, la Colaboración Cochrane, el National Institute of Clinical Excellence, la Scottish Intercollegiate Guidelines Network y publicaciones como *Clinical Evidence* o *Uptodate* (http://www.gradeworkinggroup.org/society/index.htm). En nuestro entorno, el Programa Nacional de Elaboración de Guías de Práctica Clínica del Sistema Nacional de Salud (http://www.guiasalud.es/web/guest/gpc-sns) también ha comenzado a utilizarlo (manual MSC).[13]

El sistema GRADE propone varios factores para evaluar la confianza en los resultados, de los cuales algunos pueden disminuir nuestra confianza y otros la pueden aumentar.[12] A los ensayos clínicos se les asigna de entrada una calidad alta, pero pueden ser penalizados. A los estudios observacionales se les asigna de entrada una calidad baja, pero en algunas ocasiones la confianza en ellos puede aumentar. Este sistema explícito de subida y bajada es propio y único del sistema GRADE.

La figura 3 resume los diferentes factores que pueden disminuir la confianza en la estimación del efecto observado.[14] Éstos son, fundamentalmente, que los estudios presenten limitaciones en el diseño o la ejecución (riesgo de sesgo), que los resultados no sean concordantes, que no se disponga de evidencia directa para nuestra pregunta de interés, que los resultados sean imprecisos o que se sospeche un sesgo de publicación. La ponderación global de estos factores, que limitan nuestra confianza en los resultados, determinará que la confianza aumente o disminuya. El sistema GRADE establece que la calidad global es la menor entre los desenlaces o resultados de interés. Por último, el sistema GRADE reconoce que la opinión de los expertos influye en la evaluación de la evidencia disponible (con independencia de su diseño), pero no la considera un tipo de evidencia en sí misma.

El sistema GRADE proporciona, por tanto, un marco para que tanto los periodistas como los profesionales de la salud nos orientemos sobre

la confianza que podemos depositar en los resultados de la investigación. Por lo que respecta al periodismo, estas premisas pueden resultar útiles para diferenciar el grano de la paja.

¿Qué puede disminuir la confianza en los resultados?

Problemas de diseño y ejecución

Las limitaciones en el diseño y la ejecución difieren entre los ensayos clínicos aleatorizados y los estudios observacionales. En los primeros, los factores clásicos son los relacionados con una adecuada aleatorización, un cegamiento de los participantes y un seguimiento adecuado y completo de los participantes. La aleatorización, si está correctamente realizada y con un número suficiente de sujetos, proporcionará dos poblaciones o más que tendrán un pronóstico similar, pues el azar habrá repartido de manera equilibrada (pronósticamente) a los participantes en el estudio, y con ello estaremos evitando un sesgo de selección. En el caso del cegamiento, esta medida evitará, entre otras cosas, que los pacientes asocien el hecho de recibir uno u otro tratamiento con el efecto que experimenten o evalúan (sesgo de detección), y que los profesionales sanitarios traten de manera diferencial a un grupo o a otro (sesgo de realización). En el caso del seguimiento, es crucial que no haya sujetos de quienes no tengamos información a partir de un momento dado, pues pueden tener un pronóstico diferente al de los que permanecen en el estudio y, por tanto, sesgar de manera importante los resultados (sesgo de desgaste).

En los estudios observacionales se consideran otros factores, como la presencia de unos criterios de selección de la población inapropiados, las mediciones inadecuadas para la exposición o el desenlace de interés, un mal control de los factores de confusión o un seguimiento incompleto, entre otros.

Resultados no concordantes

La calidad de la evidencia disminuye si los resultados son heterogéneos o incongruentes, es decir, si los distintos estudios muestran resultados muy diferentes. Debe valorarse, además, si tras explorar las razones que pudieran explicar las diferencias (por ejemplo, diferencias en la población, la intervención, los desenlaces de resultado o el riesgo de sesgo), éstas persisten. En caso de no identificar las razones de la heterogeneidad, la confianza disminuye porque podría haber diferencias reales entre las estimaciones del efecto proporcionadas por los estudios.

La concordancia depende muy a menudo de la contextualización con la investigación previa. Esta información debería estar en el comunicado

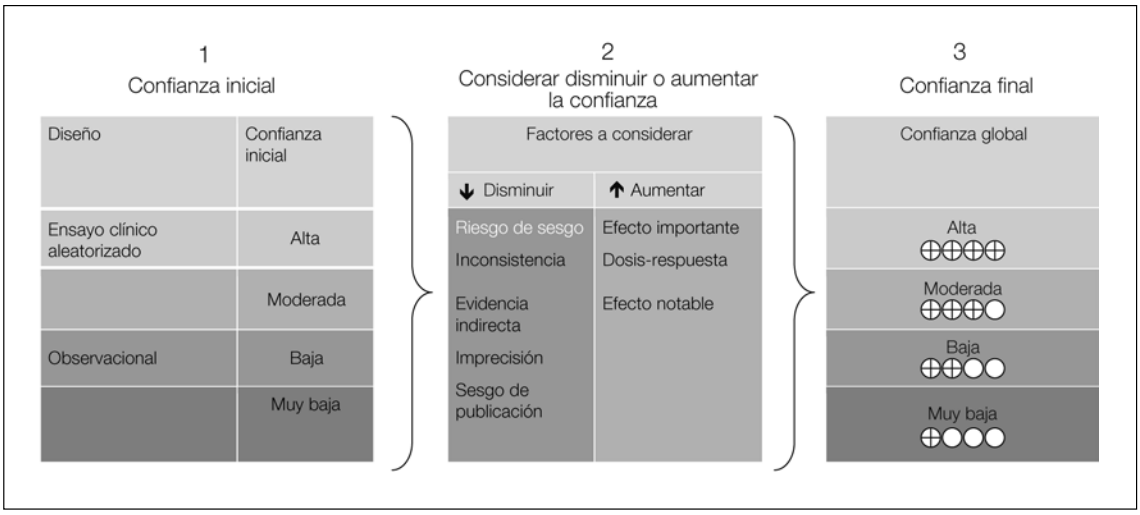


Figura 3. Evaluación de la confianza (calidad) y factores modificadores según el sistema GRADE.[14]

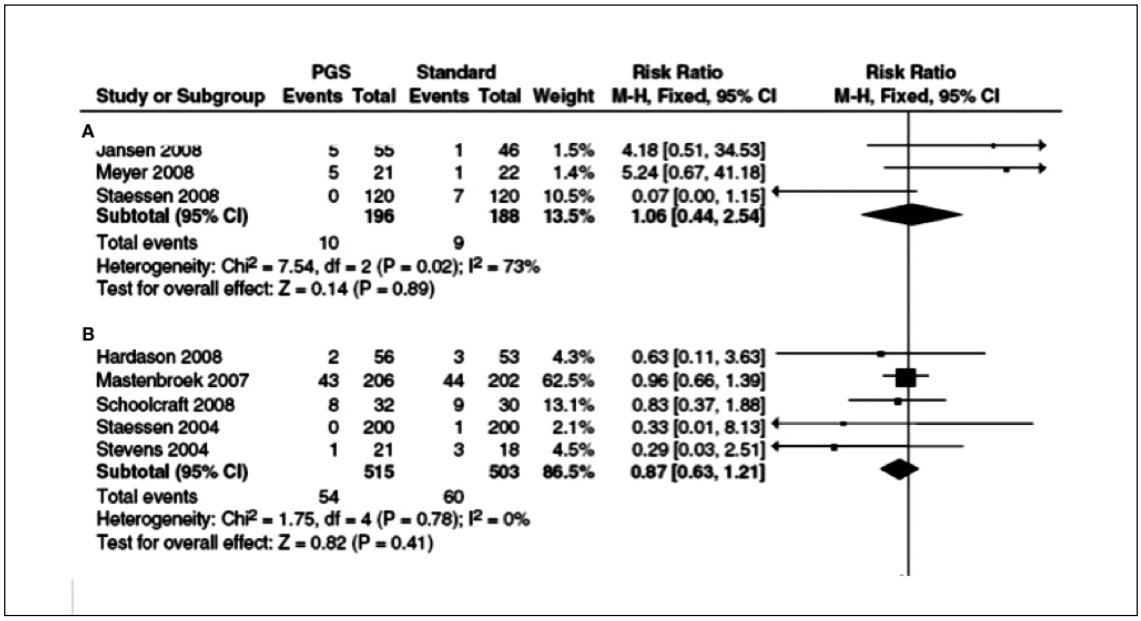


Figura 4. Revisión sistemática sobre la efectividad del cribado genético preimplantacional comparado con el tratamiento estándar en parejas sin trastornos genéticos. A) Metaanálisis con importante heterogeneidad/variabilidad. B) Metaanálisis con escasa heterogeneidad/variabilidad. PGS: *preimplantation genetic screening;* IVF: *in vitro fertilization*.[15]

de prensa, e idealmente, como comentábamos antes, mediante una revisión sistemática previa. Un estudio no es una isla y siempre hay un contexto. Si la información no está disponible, es recomendable que los informadores consulten varias fuentes independientes, incluyendo un experto en metodología y estadística, sea o no clínico. Además, a menudo puede ser interesante considerar la visión tanto de la atención primaria como de la hospitalaria. Estas fuentes nos ayudarán no sólo a evaluar la heterogeneidad, sino también el resto de los factores.

Si disponemos de una revisión sistemática, una fórmula práctica para valorar la consistencia es observar si los intervalos de confianza se solapan en el gráfico correspondiente. También se dispone de pruebas estadísticas más complejas, con las cuales no tienen por qué estar familiarizados los periodistas, sino nuestra fuente experta en métodos. En la figura 4 A observamos, en la parte superior, un primer rombo que corresponde al análisis conjunto de los ensayos.[15] Los intervalos de confianza son muy amplios y se solapan escasamente. Asimismo, los efectos son muy dispares. Estaríamos, por tanto, ante resultados poco concordantes. En cambio, el rombo de la figura 4 B corresponde al análisis conjunto de cinco estudios que presentan intervalos de confianza más precisos, con mayor solapamiento y efectos más similares. En este caso, los resultados son más congruentes y nuestra confianza no se verá mermada.

Resultados imprecisos

Cuanto más imprecisos sean los resultados de un estudio o de una revisión sistemática, menos confianza tendremos. Para poder valorar la precisión es fundamental fijarse en el intervalo de confianza (la línea horizontal que mencionábamos antes al explicar el metaanálisis), valorando que no se solape con la línea vertical del valor relativo 1, que indica que hay un riesgo similar en ambos grupos y que, por tanto, no hay diferencias entre las intervenciones. Asimismo, ante un intervalo de confianza preciso, si el número de sucesos (por ejemplo, el número total de reingresos o infartos) o de sujetos evaluados en los diferentes estudios son escasos, también debe considerarse disminuir la confianza, pues en ocasiones los estudios pequeños, por puro

azar, pueden proporcionar efectos beneficiosos muy importantes (que no son tales). Finalmente, hay que valorar si al considerar un extremo u otro del intervalo de confianza, teniendo en cuenta los riesgos e inconvenientes de la intervención, nuestra decisión cambiaría. Si es así, la confianza en el resultado disminuirá, por impreciso.

Información no directamente aplicable (indirecta)

La confianza en los resultados depende también de si el estudio o los estudios que estamos sopesando son suficientemente similares a los nuestros (población, intervención, comparación y desenlaces de interés). Por ejemplo, si los estudios son en animales, la evidencia será muy indirecta y nuestra confianza disminuirá de manera marcada. Si un estudio nos proporciona información sobre supervivencia, por ejemplo en el campo del cáncer y la quimioterapia, y nos interesa más la calidad de vida, nuestra confianza en que ésta mejore (o empeore menos), por ejemplo, en comparación con el tratamiento sintomático, será también baja. Esta analogía es igualmente válida en el caso del riesgo de fractura en las mujeres posmenopáusicas, si un estudio sólo nos proporciona datos sobre densidad mineral pero no sobre la calidad de vida. En ocasiones, cuando no hay comparaciones directas entre tratamientos, se realizan comparaciones indirectas mediante técnicas estadísticas. Sus resultados serán, por tanto, de menor confianza que los obtenidos mediante comparaciones directas.

Sospecha de sesgo de publicación

Finalmente, hay situaciones en las que se sospecha que existen estudios, principalmente con resultados negativos, que no se han publicado y que por tanto hay una posible sobrestimación del efecto. Esta posibilidad debe explorarse si nos encontramos con un conjunto de ensayos de pequeño tamaño, positivos y financiados por la industria. Al ser pequeños, por azar debería haber tanto resultados positivos como negativos. En estos casos se reduciría la confianza en la estimación de un efecto. Para detectar este posible sesgo se dispone de pruebas estadísticas o gráficas (por ejemplo el *funnel plot*) que a menudo las revisiones sistemáticas incluyen como parte de sus métodos, facilitándonos el trabajo. Aunque muchas veces es un aspecto difícil de valorar, es bueno tenerlo en cuenta.

¿Qué puede aumentar la confianza en los resultados?

Las situaciones que justificarían un aumento de nuestra confianza en los resultados de un conjunto de estudios son menos comunes y se aplican fundamentalmente en los estudios observacionales (sobre todo de cohortes y de casos y controles), siempre que no coexistan otras limitaciones de las anteriormente descritas (por ejemplo, limitaciones de diseño y ejecución). Estas situaciones son poco frecuentes, pero existen.

Asociación fuerte

Cuando los resultados de un estudio, sin otras limitaciones, muestran un efecto, protector o perjudicial, con una asociación fuerte (riesgo relativo u *odds ratio* >2 o <0,5) o muy fuerte (riesgo relativo u *odds ratio* >5 o <0,2), la confianza en los resultados aumenta. Un ejemplo es la relación que se encuentra entre la mortalidad por cualquier causa y el consumo de tabaco, que resultó ser hasta tres veces mayor en los fumadores respecto a los no fumadores en una cohorte prospectiva de médicos británicos. Así pues, la confianza en esta asociación no es baja (punto de partida de los estudios observacionales). Del mismo modo, si el efecto de la intervención es relativamente inmediato y cambia de manera radical el pronóstico de los pacientes, nuestra confianza también aumenta. Por ejemplo, no ha sido necesario hacer ensayos clínicos sobre la colocación o no de una prótesis de cadera en pacientes con artrosis grave de esta articulación. La mejora es muy importante, el efecto es bastante inmediato y el pronóstico de los pacientes es radicalmente diferente.

Gradiente dosis-respuesta

Otro factor que aumenta la confianza en la estimación de un efecto es la existencia de un claro

gradiente dosis-respuesta, ya que nos aporta una mayor certidumbre sobre una posible relación causa-efecto. Por ejemplo, se ha comprobado que el riesgo de desarrollar una enfermedad pulmonar obstructiva crónica es proporcional al consumo acumulado de tabaco, y que es 2,6 veces mayor en los fumadores de 15 a 30 paquetes al año y 5,1 veces mayor en los fumadores de más de 30 paquetes al año.[16] La existencia de este gradiente de asociación entre el factor estudiado y el efecto aumenta la confianza en la relación entre el tabaco y la enfermedad pulmonar obstructiva crónica.

Conclusiones

La calidad es la confianza que tenemos en que los resultados de la investigación sean ciertos. Las revisiones sistemáticas son clave para contextualizar y conocer con mayor seguridad el efecto de las intervenciones, pues nos proporcionan todos los estudios disponibles para una determinada cuestión.

El sistema GRADE aporta un marco explícito y riguroso para evaluar la calidad que podemos depositar en los resultados de la investigación. La calidad (o la confianza) no sólo está determinada por el diseño de los estudios. La confianza puede disminuir debido a problemas de diseño y ejecución, resultados imprecisos o no concordantes, evidencia indirecta o sospecha de sesgo de publicación. La confianza puede aumentar, en los estudios observacionales, cuando hay un gradiente dosis-repuesta o si el efecto observado es muy importante (y no se observan otras limitaciones).

Bibliografía

1. Systems to rate the strength of scientific evidence. Summary, evidence report/technology assessment: number 47. AHRQ Publication No. 02-E015. Rockville, MD: Agency for Healthcare Research and Quality; 2002.
2. Laine C, Taichman DB, Mulrow C. Trustworthy clinical guidelines. Ann Intern Med. 2011;154:774-5.
3. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al.; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924-6.
4. The GRADE Working Group. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches. BMC Health Serv Res. 2004;4:38.
5. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: síntesis of best evidence for clinical decisions. Ann Intern Med. 1997;126:376-80.
6. Gisbert JP, Bonfill X. ¿Cómo realizar, evaluar y utilizar revisiones sistemáticas y metaanálisis? Gastroenterol Hepatol. 2004;27:129-49.
7. Cochrane AL. Effectiveness and efficiency. Random reflections on health services. London: Nuffield Provincial Hospitals Trust; 1972.
8. Bero L, Rennie D. The Cochrane Collaboration. Preparing, maintaining, and disseminating systematic reviews of the effects of health care. JAMA. 1995;274:1935-8.
9. Higgins JPT, Green S, editores. Cochrane handbook for systematic reviews of interventions, version 5.1.0. Actualizado en marzo de 2011. The Cochrane Collaboration, 2011. Disponible en: www.cochrane-handbook.org
10. Crowley P, Chalmers I, Keirse MJ. The effects of corticosteroid administration before preterm delivery: an overview of the evidence from controlled trials. Br J Obstet Gynaecol. 1990;97:11-25.
11. Clarke M, Hopewell S, Chalmer I. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. Lancet. 2010;376:20.
12. Alonso-Coello P, Rigau D, Solà I, Martínez García L. La formulación de recomendaciones en salud: el sistema GRADE. Med Clin (Barc). 2013;140:366-73.
13. Grupo de Trabajo sobre GPC. Elaboración de guías de práctica clínica en el Sistema Nacional de Salud. Manual metodológico. Guías de práctica clínica en el SNS: I+CS. N8 2006/0I. Madrid: Plan Nacional para el SNS del MSC. Instituto Aragonés de Ciencias de la Salud-I+CS; 2007.
14. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: introduction – GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64:383-94.
15. Checa MA, Alonso-Coello P, Solà I, Robles A, Carreras R, Balasch J. IVF/ICSI with or without preimplantation genetic screening for aneuploidy in couples without genetic disorders: a systematic review and meta-analysis. J Assist Reprod Genet. 2009;26(5):273-83.
16. Miravitlles M, Soriano JB, García-Río F, Muñoz L, Durán-Taulería E, Sánchez G, et al. Prevalence of COPD in Spain: impact of undiagnosed COPD on quality of life and daily life activities. Thorax. 2009;64:863-8.

El intervalo de confianza se interpreta mejor que el valor de *p*

Cobo / Ventura

Diálogo 1
Herramientas estadísticas, buenos consejos y cierta intuición periodística

Ainhoa Iriberri

Los periodistas que acudimos a la jornada de bioestadística organizada el pasado 14 de febrero por la Asociación Española de Comunicación Científica y la Fundación Dr. Antonio Esteve pedíamos algo imposible: que en menos de 12 horas los expertos nos enseñaran a los comunicadores a informar bien sobre un estudio publicado en una revista científica. Se trataba de aprender a evitar algunos de los grandes males del periodismo científico: la exageración, la falta de rigor y la descontextualización, entre otros.

Así, estos conceptos salieron a la luz en el primer diálogo entre periodistas y bioestadísticos, que sirvió también como punto de encuentro. Porque ni los bioestadísticos son esas personas que, con sus enrevesados términos, dificultan muchísimo la comprensión de un estudio, ni los periodistas los irresponsables que algunos de los primeros pueden pensar. Informamos lo mejor que podemos con nuestros conocimientos y las opiniones de expertos acreditados.

Quizás el error, como quedó de manifiesto en la jornada, es que la bioestadística es una disciplina nueva y algunos investigadores –y la mayoría de los periodistas– carecen muchas veces de conocimientos suficientes sobre la materia. Pero el diálogo entre ambos agentes también demostró que compartimos un deseo común: informar con rigor de los avances científicos.

Pirámides y confianza

Los periodistas estamos acostumbrados a trabajar con pirámides. Desde la famosa pirámide invertida, o regla de las «5 w», sabemos que la información hay que clasificarla y jerarquizarla.

Por ello no nos extrañó cuando se habló del concepto de «pirámide de la evidencia». No todos los estudios científicos son iguales, y no debemos confiar en todos ellos por igual. De nuevo, hemos de recurrir a una pirámide; en este caso, la de la evidencia.

La pirámide ilustra que hay una jerarquía de la confianza que nos merecen los distintos tipos de estudios, y el interés de un grupo de trabajo que comenzó en 2000, el grupo GRADE (siglas en inglés de Clasificación de la Evaluación, Desarrollo y Valoración de las Recomendaciones). El objetivo de este colectivo no era otro que desarrollar un sistema común y razonable para calificar la calidad de la evidencia y la fuerza de las recomendaciones. Un sistema que, obviamente, puede servir también a los periodistas a la hora de elaborar sus informaciones sobre un avance científico.

En la base de la pirámide están los estudios experimentales realizados con animales de laboratorio, luego se encuentran los distintos tipos de estudios observacionales, y en la cúspide los mejores ensayos clínicos y las revisiones sistemáticas (con o sin su correspondiente metaanálisis).

A partir de ahí, podríamos pensar que está todo dicho. A la hora de valorar si informamos sobre un estudio, sólo hemos de situarlo en la pirámide. Si la confianza es baja, no se informará sobre él; si es alta, intentaremos que vaya en portada. Esto que dicta la lógica está, sin embargo, muy lejos de la realidad.

Como los ponentes dejaron muy claro desde el principio, tal jerarquía es flexible, como lo es la clasificación de la confianza que nos ofrece. Estudios observacionales que a priori merecen poca confianza pueden aumentarla si reunen una serie de requisitos, que la mejoran. Del mismo modo, estimaciones procedentes de un ensayo clínico, incluso aleatorizado –lo que es clave para empezar a confiar en este tipo de estudio– pueden no ser buenas si el ensayo está mal diseñado y ejecutado.

Sin embargo, la confianza en los resultados no es el único criterio que los periodistas seguimos a la hora de informar. Obviamente, el primero será si la información en cuestión nos parece interesante, es decir, de interés periodístico. Otros factores, como la revista científica en que se publica un determinado estudio, también pesarán en nuestra decisión, ya que no es lo mismo que un trabajo aparezca en una revista de alto impacto o en una de tercera fila. Para saberlo, hay un *ranking* que clasifica las revistas por su factor de impacto (citas recibidas), el *Journal Citation Reports* del Institute for Scientific Information, actualmente Thomson Reuters, por la empresa que lo elabora. Un obstáculo para los periodistas: su acceso es de pago, aunque muchas instituciones, entre ellas la Fundación Española para la Ciencia y la Tecnología, lo ponen a disposición de los investigadores. Una razón añadida para que se cuente entre nuestras fuentes más consultadas.

Revisión frente a metaanálisis

Uno de los primeros conceptos que se discutieron en este diálogo fue la diferencia entre revisión sistemática y metaanálisis. Las revisiones sistemáticas (con o sin metaanálisis) pueden cambiar la práctica clínica y, por tanto, son más que dignas de merecer un hueco destacado en los periódicos. Pero como explicaron los especialistas, no todas las revisiones pueden tener metaanálisis. ¿Por qué? Lo primero que hay que tener claro es que las revisiones “narrativas” del pasado han dado lugar a las revisiones sistemáticas, con una metodología explícita, idealmente especificada en un protocolo. Una vez obtenidos los estudios pertinentes, los incluidos en la revisión en ocasiones pueden metaanalizarse de manera conjunta, pero no siempre, por diversas circunstancias. Asimismo, los estudios pueden no haber medido una variable de manera similar, o haber reportado los resultados numéricos de distintas maneras o incluso de forma incompleta. Una cosa está clara: siempre que sea razonable, los autores llevarán a cabo un metaanálisis, algo que suma valor a la investigación porque aumenta la precisión de los resultados.

En la definición de «metaanálisis» hubo algo de controversia entre los participantes en la jornada. Según se explicó, este tipo de estudio

supone poder combinar pacientes de diferentes trabajos previos. ¿Se trata, entonces, de un artificio estadístico? Algunos de los ponentes prefirieron cambiar esa palabra (artificio) por una más neutra: método. Su argumento fue que proporciona respuesta a dos preguntas cruciales: 1) cuál es el efecto de la intervención en estudio y 2) cuál es el grado de homogeneidad de este efecto entre los diferentes estudios.

En cualquier caso, los expertos nos ofrecieron algunas claves para detectar una buena revisión sistemática. En primer lugar, la revisión de los trabajos ha de ser sistemática, es decir, se ha de buscar todo lo que se ha publicado sobre una determinada hipótesis. «Nos debe alertar que alguien realice un metaanálisis sin haber hecho una revisión sistemática; si no, no hay manera de trazar lo que hay metido ahí dentro, puede que sean los estudios favoritos de los autores. Es la principal cautela que hay que tener a la hora de mirar una revisión», señaló Pablo Alonso.

Al hablar de homogeneidad de estudios para poder incluirlos en un metaanálisis, tampoco hay que exagerar. No se trata de que sean trabajos exactamente iguales, pero sí tienen que intentar responder a la misma pregunta. Por ejemplo, si el metaanálisis evalúa el valor de una intervención terapéutica, ésta ha de ser similar en todos los trabajos, pudiendo variar la dosis, el comparador (por ejemplo, incluyendo placebo y no intervención) e incluso las características de los pacientes, aunque sólo ligeramente. Como dijo Pablo Alonso, «tiene que tener un sentido biológico y clínico a la hora de combinar los estudios, que no chirríe al incluir pacientes o intervenciones completamente diferentes; por eso hay unos criterios de inclusión y exclusión consensuados en las revisiones sistemáticas».

Los periodistas insistían: ¿cómo distinguir qué metaanálisis es el mejor? De nuevo, la clave está en la calidad. «El drama del metaanálisis es la heterogeneidad», dijo Erik Cobo.

Al final se consiguió la receta que los comunicadores estábamos buscando. Cuando se busca la respuesta a una pregunta científica polémica, una buena estrategia es averiguar, en primer lugar, si existe una revisión de la Colaboración Cochrane. Ésta es una entidad en la que colaboran más de 28.000 investigadores de más de 100 países (incluido España), y que se dedica a hacer revisiones sistemáticas sobre los temas más variados. Si hay una revisión de este tipo sobre un tema polémico, sus conclusiones son dignas de figurar en la información periodística.

Tenemos así una herramienta que los periodistas científicos hemos de consultar con regularidad, ante informaciones que llegan a la redacción, el estudio de radio o el plató de televisión. Si entramos en la web de la Colaboración Cochrane (www.cochrane.org) y accedemos a *Top 50 Reviews,* veremos algunos ejemplos interesantes, como revisiones sobre si el uso de estatinas previene la enfermedad cardiovascular en las personas sanas, si los suplementos con antioxidantes previenen la mortalidad o si la vitamina C es útil para prevenir los tan frecuentes catarros. Seguro que, como periodistas, hemos recibido informaciones sobre estos temas a lo largo de nuestra carrera; la próxima vez, sabremos qué consultar antes de decidir incluirlas o no. Es aconsejable mirar otras revisiones sistemáticas, pues hay vida más allá de la Cochrane. Una buena fuente es el metabuscador TripDatabase (http://www.tripdatabase.com/), que clasifica los resultados en diferentes categorías.

Sin embargo, está claro que los bioestadísticos no son partidarios de las soluciones fáciles. Así, aunque la Cochrane goza de su confianza, Pablo Alonso indicó que «no todas sus revisiones son de igual calidad». Pero la realidad, señalaron, es que hay estudios empíricos que comparan revisiones de esta organización y otras, y concluyen que, en términos generales, las primeras tienen una mayor calidad.

Una revisión sistemática debería tener un protocolo, hacer una pregunta específica y llevar a cabo una búsqueda rigurosa, bien descrita en la publicación. Ejemplos de buenas prácticas son las que consultan más de una base de datos, expertos, referencias bibliográficas, registros de ensayos clínicos e incluso inquieren a la industria farmacéutica. Además, han de hacer una evaluación del riesgo de sesgo, es decir, evaluar de alguna manera la calidad de los estudios

que incluyen. Si queremos profundizar más en la interpretación de una revisión sistemática, es aconsejable consultar una fuente con experiencia. Al fin y al cabo, la metodología de las revisiones sistemáticas, y la estadística en general, es toda una ciencia.

Que la bioestadística sea una ciencia no quiere decir que los periodistas no nos podamos formar en ella. Así, Erik Cobo fue más allá y recomendó a los comunicadores presentes la lectura y el estudio de la guía Prisma,[1] elaborada para evitar la inercia de los investigadores y consistente en una serie de ítems que éstos deben preguntarse para saber si están elaborando correctamente un metaanálisis. Por ejemplo, todo buen metaanálisis ha de incluir información sobre cuándo se empezaron a recoger los datos y cuando cesó su recopilación. Es útil también, por tanto, para evaluar la calidad de los metaanálisis ajenos y no sólo los que se están llevando a cabo.

Erik Cobo recordó por qué se creó esta guía y otras similares, que pueden consultarse en la web de la Red Equator (http://www.equator-network.org): «Se supone que los investigadores son los profesionales que rompen fronteras, que consiguen que la sociedad vaya más lejos, pero hay casos clarísimos de cosas absurdas que se hacían en todas las revistas de investigación sobre las que los estadísticos llevábamos años advirtiendo». Antes de que surgieran este tipo de iniciativas, no era raro que los investigadores utilizaran como argumento para emplear una determinada metodología que era la que se utilizaba en una revista de referencia. Por eso, los metodólogos se unieron con placer a la iniciativa para buscar transparencia en las investigaciones científicas, iniciada por el grupo de investigadores y editores médicos herederos del llamado Grupo de Vancouver, que lideró la mejora de la calidad de los originales científicos. ¿Era esto necesario? «Hay una argucia que se utiliza mucho: un investigador pide al estadístico que le haga el análisis lo más complicado posible, para que revisores y editores no entiendan nada y no les quede más remedio que tragar; esto se usa y es contrario a las ideas originales de los estadísticos", apuntó Erik Cobo.

La p y el intervalo de confianza

El bioestadístico añadió algunas afirmaciones sin duda curiosas. Por ejemplo, que p (ese número que establece hasta qué punto una hipótesis puede ser debida a la casualidad) «ha de jubilarse» y que es mejor fijarse en otro parámetro, como es el intervalo de confianza. «Los investigadores, al publicar, siguen la tradición. Por eso, esta jubilación sólo se conseguirá si disponemos de guías sobre cómo publicar artículos. No olvidemos que las revistas científicas no publican artículos buenos, sino los mejores que les llegan», afirmó.

Los periodistas presentes en la jornada preguntamos por ejemplos prácticos. Al fin y al cabo, teníamos que salir de esta reunión sabiendo cazar al vuelo un metaanálisis y diferenciarlo, además, de otros tipos de estudios.

El ejemplo clásico, señaló Pablo Alonso, es una revisión sistemática publicada en la década de 1990 que evaluaba la eficacia de un fármaco, la lidocaína, para prevenir arritmias en los pacientes que habían sufrido un infarto agudo de miocardio. El medicamento se aplicaba desde 20 años atrás. Algunos estudios observacionales parecían demostrar un posible efecto beneficioso en la disminución del riesgo de muerte, pero algunos investigadores empezaron a cuestionar la idea y se iniciaron ensayos aleatorizados, que compararon la evolución de los pacientes tratados con lidocaína con la de aquellos que no recibían el fármaco. Aunque estos estudios por separado apuntaban (de manera no concluyente) que los enfermos que tomaban el medicamento morían más, se continuaron haciendo ensayos clínicos y hubo que esperar a que alguien realizará una revisión sistemática (con metaanálisis) para que cambiara la práctica clínica y se abandonara el uso de lidocaína con este fin. Dicho metaanálisis se publicó en la década de 1990, pero si se hubieran metaanalizado los estudios llevados a cabo en la década de 1980 ya se hubiera visto que el efecto era claramente perjudicial. Conclusión: durante muchos años, miles de personas fueron tratadas innecesariamente y con consecuencias muy graves. ¡Como para no valorar las revisiones sistemáticas!

Estudios observacionales

Cuando parecía que ya todo estaba explicado, el epidemiólogo José Luis Peñalvo, del Centro Nacional de Investigaciones Cardiovasculares, añadió un nuevo dato en el cual fijarse: que no es lo mismo un metaanálisis de estudios experimentales que uno de trabajos observacionales.

Los estudios observacionales ocupan muchas veces los distintos medios de comunicación. Cada vez que leemos informaciones que asocian un determinado factor de riesgo a un efecto, es más que probable que la información venga de un trabajo de este tipo, por lo que conviene saber un poco más sobre ellos. A fin de cuentas, como observó Cobo, la búsqueda de causas es la pasión favorita de las mentes más racionalistas, a diferencia de la confirmación de efectos que persiguen las más pragmáticas y empíricas.

La primera clave sobre los estudios observacionales es básica para el periodismo. Con los estudios observacionales no puede inferirse causalidad. «Lo único que podemos deducir es una asociación o una relación y, dentro de ésta, todos los grados de gris: fuerte, modesta, no significativa, una tendencia…», señaló Peñalvo. En la escala de fuerza de asociación, hay tres tipos de diseño: de casos y controles, transversales y de cohortes.

Los estudios de casos y controles responden a un tipo de diseño observacional que es muy atractivo, porque es muy rápido. No hace falta mucha infraestructura y son perfectos para estudiar hipótesis sobre enfermedades de baja prevalencia. ¿Cómo se llevan a cabo? Se acude a un hospital y se seleccionan todas las personas que ingresen con una determinada enfermedad, a las que se denominará «casos». Después se establecen una serie de criterios para escoger unos «controles», que estarán emparejados por tener similitudes con los casos. Posteriormente, se utilizan registros o se hacen preguntas a los participantes, por ejemplo sobre elementos a los que están expuestos, y se establecen asociaciones. Son, en definitiva, estudios rápidos y fáciles de hacer que, en ocasiones, se realizan dentro (anidados) de estudios de cohortes.

Las cohortes tienen, por definición, una duración larga. El ejemplo clásico es la cohorte del estudio Framingham, un trabajó que empezó en 1948 en el pueblo de la costa este de Estados Unidos del mismo nombre y que sirvió para establecer los factores clásicos de riesgo cardiovascular. Actualmente se está estudiando la tercera generación, los nietos de los primeros voluntarios que accedieron a que un equipo de cardiólogos les vigilara de por vida. Ahora, los factores que se estudian están más centrados en la herencia.

Por último, son importantes los estudios observacionales transversales, también muy habituales y muy presentes en los medios de comunicación. Un ejemplo claro son las encuestas nacionales de salud (en España se realizan cada 2 años). Se trata de establecer tendencias con datos a partir de estudios observacionales. Un ejemplo de este tipo de estudio son las informaciones que, a partir de combinar la información de varios “cortes” transversales en la población, relacionan la entrada en vigor de la Ley Antitabaco con la disminución de episodios cardiovasculares o de muertes por esta causa.

La importancia de la causalidad

En realidad, destacaron los expertos, nunca podrá decirse con total seguridad que una circunstancia ha llevado a la otra, y como ocurre siempre con la epidemiología observacional, la relación no necesariamente será causal. El debate se complicó aún más cuando Erik Cobo reflexionó en alto: «¿Y si, en ocasiones, no importara realmente si es causal o no?». Así, este estadístico habló de un ejemplo concreto: los estudios que asocian las modificaciones de las leyes de seguridad vial con la reducción del número de muertos en la carretera. «Si decidimos titular *La nueva Ley ha evitado 100 muertes* no lo hacemos bien; si titulamos *Desde la entrada en vigor de la Ley ha descendido el número de muertes,* el titular es neutro y correcto», explica Cobo, que sin embargo añadió: «A mí, como ciudadano, los dos titulares me son útiles. A veces los estadísticos nos ponemos muy antipáticos», concluyó.

De este primer diálogo entre periodistas y estadísticos queda claro que existen muchas herramientas que los primeros podemos utilizar para informar correctamente sobre las investigaciones científicas. No obstante, es evidente que sigue haciendo falta formación, una cierta intuición y, sobre todo, consultar a los expertos. Sólo que ahora, en la agenda, habrá que añadir un apartado al de investigadores, cardiólogos, físicos o cualesquiera categorías que utilicemos habitualmente: los bioestadísticos, que sin duda tienen mucho que decir.

Bibliografía

1. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6:e1000097. Disponible en: http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000097

Una cosa es conocer la causa y otra dar con la solución

Casino / Ventura

Diálogo 2
Sobre los estudios observacionales y su tratamiento periodístico

Pablo Francescutti

El segundo diálogo se centró en los estudios observacionales y su tratamiento en los medios de comunicación, pero desbordó este marco inicial para adentrarse en otras cuestiones de interés periodístico. La sesión, también moderada por Gonzalo Casino, contó con la presencia de los dos expertos de la mañana, Erik Cobo y Pablo Alonso, y dos periodistas, Esperanza García Molina y Pablo Francescutti.

Esperanza García Molina inició el debate con una cuestión muy concreta y de gran valor práctico en términos periodísticos: ¿en qué elementos debe fijarse un periodista que tiene entre manos un estudio observacional para saber si merece difusión?

Alonso: Nosotros no podemos entrar a valorar la importancia periodística de un estudio, porque es algo que os toca a vosotros determinar. Lo que sí podemos es repasar lo que se ha dicho a lo largo de la jornada: diferentes diseños ofrecen diferentes niveles de confianza de entrada, confianza alta los ensayos clínicos y baja los observacionales. Ése es un primer criterio. Otros parámetros a tener en cuenta son que la cuestión que se plantearon los investigadores sea relevante y que el diseño elegido sea adecuado para responder a esa cuestión. Conviene tener presente que en cada estudio observacional, al igual que en los ensayos clínicos, hay una serie de posibles sesgos, sesgos de detección, de realización y otros. Dependiendo del estudio habrá que hacer una serie de preguntas para saber si ha sido bien ejecutado, cómo se reclutó la muestra, si se ajustaron los factores de confusión, y cómo se hizo el seguimiento. Cada estudio tiene unas

características sobre las que hay que hacerse preguntas específicas, y eso requiere un conocimiento básico de cómo leer críticamente un estudio observacional (ya sea, por ejemplo, de casos y controles o de cohortes), de cómo valorar su diseño y ejecución (riesgo de sesgo), y de qué factores aumentan o disminuyen la confianza, por ejemplo, sabiendo que ésta será mayor si se ha establecido una asociación muy importante o existe un gradiente dosis-respuesta (en el caso de los observacionales), y que será menor si los resultados son inconsistentes, si son imprecisos y si no son directamente aplicables a la población que interesa.

García Molina: Mi segunda pregunta tiene que ver con la causalidad. ¿Nunca puede decirse de un estudio observacional que ha establecido una relación causa-efecto?

Cobo: Para que nos entendamos, siempre es delicado hablar de causalidad, incluso en un ensayo clínico. Recordemos que existen dos formas de razonar: una es la deductiva, en la que partimos de unos principios que damos por ciertos, que podemos llamar premisas; y la otra es la inductiva, en la que partiendo de unos pequeños casos intentamos generalizar lo observado a muchos otros. Decía esta mañana que no tenemos que hablar de las leyes de Newton sino del modelo de Newton, porque las piedras no están obligadas a seguirlo o ir a la cárcel si no lo hacen. Por eso resulta una vanidad decir que estamos escribiendo las leyes de la naturaleza; lo que tiene mucho mérito es que somos capaces de elaborar modelos que nos permiten construir edificios que se aguantan, y eso ya es mucho. Un modelo es una esquematización de la realidad en un lenguaje muy abstracto, que tiene la ventaja de que lo interpretan igual un japonés y un americano. Pero estos modelos también tienen sus propias premisas. Todos los ensayos clínicos, que se sitúan en el nivel máximo de evidencia, tienen premisas; por ejemplo, el efecto que observamos en una persona es independiente del efecto que observamos en otra. Esta premisa puede no cumplirse en el caso de una vacuna, porque si un individuo no se contagia, será más difícil que otro lo haga; por eso algunos ensayos se diseñan para poder estudiar ese contagio entre unidades. Pero por lo habitual se asume que las unidades se comportan independientemente y se hace el análisis sobre esta base. Por lo tanto, la lectura correcta de un ensayo clínico sería la siguiente: asumiendo ciertas todas las premisas, entre las que resalta que el efecto es independiente en las unidades –y habiéndose cumplido todo lo recién dicho por Pablo Alonso– podemos afirmar que el efecto de esta intervención es el siguiente: "…". Por eso, en general debemos ser muy modestos al hablar de los resultados, incluso delante del mejor diseño que tenemos, que es el ensayo clínico. En lo relativo a los estudios observacionales, tenemos que ser muy transparentes con lo que hemos observado y lo que interpretamos. Faltaría más que los autores del estudio Framingham no terminasen postulando que, si se logra bajar la presión arterial de las personas, a largo plazo bajarán los eventos cardiovasculares. Es magnífico que interpretaran así sus resultados, pero tienen que dejar claro que lo que identificaron es una "simple" relación entre dos variables: aquellos con la presión más alta fueron luego quienes sufrieron más eventos cardiovasculares al cabo de unos años. Luego, por supuesto, pueden tocar en la discusión aspectos que no fueron observados en el estudio, y plantear la hipótesis de que si se consigue bajar la presión arterial de las personas, quizás disminuyan a largo plazo los ictus, los infartos, etcétera. Este razonamiento especulativo es lícito, pero hay que dejar claro que lo estás haciendo. Volviendo a tu pregunta: ¿pueden estos autores hablar de causalidad? Sí, pero dejando muy claro que esa relación no se deduce directamente de ningún estudio si no le añade alguna premisa más. Y por supuesto, los observacionales requieren más premisas. Una diferencia del ensayo clínico con el observacional es que cuando le dices a una persona que haga algo, lo hará o no lo hará. Si a alguien le dices que no se vaya de su comunidad durante su tiempo de vida porque quieres hacerle un seguimiento durante 30 años, no tiene por qué hacerte caso. Estas impurezas puedes verlas en un estudio experimental, pero en un estudio observacional no, por lo que es una incertidumbre adicional que conviene destacar.

Casino: Quiero reformular la pregunta de García Molina, que es clave: ¿cuándo puede hacerse una inferencia causal? Pongo un ejemplo: la relación entre el tabaco y el cáncer. Hace unos 60 años se puso en marcha un célebre estudio observacional: el de los médicos británicos, que se prestaron a participar distribuyéndose en dos cohortes, una de fumadores y otra de no fumadores, para ver qué consecuencias tenía para su salud. Recordemos que por aquel entonces incluso algunos médicos aparecían fumando en los anuncios de cigarrillos. Fijaos cómo ha cambiado la situación desde mediados del siglo XX, cuando los facultativos se mostraban fumando, hasta el día de hoy en que se nos dice «Fumar mata» o «Fumar produce cáncer de pulmón». Todo empezó con un estudio observacional, a principios de la década de 1950. En 1964, la Food and Drugs Administration de Estados Unidos ya asociaba el tabaquismo con un mayor riesgo de cáncer de pulmón. Con el tiempo y la acumulación de pruebas científicas, se ha pasado de establecer una asociación a establecer una relación causal.

Cobo: Es magnífico tu ejemplo. Aquí se cumplen de manera impecable y perfecta los criterios fijados por Bradford Hill: si se cumplen, la interpretación más plausible de la relación entre tabaco y cáncer de pulmón es la causalidad. La prueba definitiva la tendríamos el día en que hiciéramos un diseño experimental asignando niños y adolescentes a grupos de fumadores y no fumadores. Pero eso, afortunadamente, no lo podemos hacer; no podemos decir «te ha tocado el grupo fumador de los 13 a los 43 años». Si alguien me pregunta si disponemos de evidencia suficiente para poner en las cajetillas de cigarrillos la leyenda «Fumar mata», tendré que decir que no tenemos el mismo nivel de evidencia que se exige para introducir un fármaco en el mercado, ya que nos basamos en bastantes más presunciones. Dicho esto, añadiré que ha sido triste que hayamos esperado 30 años para colocar dicha leyenda en las cajetillas, porque se tenía que haber hecho caso a Hill mucho antes y, en consecuencia, haber puesto ese anuncio quizás en la década de 1960. La pregunta que debemos formularnos no es «¿qué sabemos sobre el tabaco?» sino «¿con qué decisión viviremos mejor?». Sanders Greenland, el gran epidemiólogo, zanjó la discusión en los años 1990 diciendo que el problema no radica en lo que no sabemos, sino en decidir qué es lo más sensato hacer. ¿Qué daño causaremos a la humanidad si alertamos de que el tabaquismo mata y luego resulta que la causa del cáncer es, por ejemplo, cierto contaminante común en todas las fábricas? Que algunos dejarán de fumar y unas empresas dejarán de ganar dinero. ¿Y qué daño causaremos si no ponemos el anuncio y más tarde se confirma que el tabaco es realmente la causa? Los epidemiólogos tienen la respuesta: 60.000 muertes al año en España.

Francescutti: Me gustaría profundizar en el asunto planteado por mi colega Esperanza García Molina, esto es, los criterios que debemos seguir los periodistas a la hora de interpretar datos biomédicos. Quiero referirme a la importancia del tamaño de las muestras de los estudios observacionales, y para ello haré un paralelismo con los sondeos electorales. A los periodistas nos han dicho que una muestra de entrevistados inferior a 1000 personas no es representativa de la opinión de toda la sociedad. ¿Podemos trasladar ese parámetro a los estudios observacionales y clínicos? ¿Cuánta gente debe participar para fiarnos de sus resultados? ¿Es el tamaño de la muestra el criterio decisivo, o debemos prestar más atención a la fuerza de la asociación o causalidad identificada?

Cobo: El tema que planteas exige una respuesta técnica. Para seguir en el plano de los sondeos, remitámonos a la opinión de los catalanes sobre el proceso soberanista. Imaginemos que en mi bolsillo derecho llevo la opinión de 5000 amigos míos, y en el izquierdo la de 100 catalanes escogidos al azar por toda Cataluña. ¿Cuál preferís? Obviamente la opinión de los 100 escogidos al azar, porque más importante que la cantidad de los datos es su calidad. La estadística mide el azar; no puede medir cómo son mis amigos, pero sí cuantificar hasta qué punto el azar ha influido en que sean de un color político u otro. La respuesta está en el intervalo de confianza. El mérito de la estadística es el de precisar el intervalo de con-

fianza. Y eso autoriza a decir que las opiniones expresadas por estos 100 individuos nos merecen una confianza del 90%, pongamos por caso. Y si hubieran sido 2000 encuestados, la confianza sería mucho mayor. Ciertamente, en las encuestas hechas al azar intervienen otros factores: que todos los votantes respondan, que no mientan, que no cambien de intención de voto en los días siguientes… Pero si se cumplen esas premisas, la estadística sabe cuantificar el margen de error y el intervalo de confianza. Los encuestadores honestos te dirán que hubo un 15% de entrevistados que no quisieron decir lo que votaban, y entonces lo más honrado sería estimar hasta qué punto esas "no respuestas" afectan sus conclusiones. En el *blog* de la Sociedad Catalana de Estadística se analiza el patinazo generalizado de las encuestas en las últimas elecciones autonómicas, como también lo hubo en 2003. ¿Cómo interpretar estos fallos? Asumiendo que las encuestas no pueden predecir el cambio de voto en las dos o tres semanas previas a las elecciones, un lapso en el que en Cataluña sucedieron muchas cosas que podían cambiar el voto. Intuyo que esas encuestas no son capaces de detectar los cambios finales, pero si tomamos las encuestas a escala global, lo que parece mentira es lo mucho que aciertan preguntando a tan pocos.

Alonso: Aparte del intervalo de confianza, muy útil para determinar si los resultados son precisos o no, existe una regla de la abuela: cuando los datos proceden de ensayos clínicos con menos de 100 eventos, la confianza que nos merecen es baja o muy baja; cuando la muestra cuenta con cientos de eventos, nos inspira una confianza moderada; y cuando disponemos de miles de pacientes o eventos, una confianza alta.

García Molina: He visto publicados estudios neurológicos llevados a cabo con resonancia magnética que sacan conclusiones con apenas 10 pacientes.

Casino: Está claro que cuando nos están hablando de unas pocas decenas de pacientes, hay que extremar las cautelas y tener cuidado con las conclusiones.

Cobo: Un matiz: cuando se describió el sida, ¿en cuántos casos se basaron? En muy pocos. Si se trata de un estudio muy novedoso, unos pocos casos podrán bastar para sacar conclusiones de interés; pero si quieres aportar algo nuevo sobre un tema conocido, debes seguir las reglas que apunta Alonso.

Alonso: Es verdad. Tomemos el descubrimiento del valor terapéutico de la adrenalina en el *shock* anafiláctico. El cambio que produjo en su tratamiento fue tan espectacular que no se necesitaron muchos ensayos para darlo por bueno.

Francescutti: Algo similar ocurrió cuando descubrieron que un tratamiento antibiótico curaba definitivamente la úlcera estomacal.

Alonso: Sí, porque se produjo un efecto muy llamativo. De todos modos, mantengamos la prudencia ante estudios con un intervalo de confianza estrecho, en los que con muy pocos eventos y muy pocos pacientes puede observarse un efecto importante. Hay trabajos que hablan de la fragilidad del valor p en tales estudios, ya que cambiando el desenlace *(outcome)* de uno o dos casos tales efectos desaparecen por completo, de modo que extremad la precaución.

Francescutti: Quiero hacer una reflexión acerca de la pirámide de la evidencia que vimos esta mañana. Cuando la explicabais caí en la cuenta de que en periodismo se suele invertir esa pirámide: las noticias basadas en estudios in vitro y con ratones, situadas en el nivel más bajo de confianza, compiten por la primera plana con los estudios epidemiológicos, más fiables. Los metaanálisis prácticamente no son noticia, salvo en revistas médicas o reportajes. Me parece lógico que una publicación especializada destaque un estudio con células o ratones, pero me pregunto cuál es el justo lugar que debemos dar a sus resultados cuando escribimos para un medio generalista.

Casino: Voy a devolver la pelota a los científicos. Que los periodistas hablemos de estudios en ratones es a menudo una consecuencia de

la influencia de la maquinaria informativa de la que forman parte los centros de investigación, las revistas de impacto, los comunicados de prensa y los investigadores. No veo el problema sólo en los periodistas que se hacen eco de esos estudios preliminares, sino también en los científicos que se prestan a salir en la radio y otros medios de comunicación para airear, por ejemplo, los efectos beneficiosos del aceite de oliva en animales, sobre todo cuando el oyente escucha la parte final de sus declaraciones sin saber si están hablando de ratones o personas. En muchos casos, la llamada de atención debería dirigirse a los investigadores que quieren salir en los medios para promocionar su carrera profesional y se prestan a informar de temas de salud a partir de estudios en ratones. En cuanto a la pregunta de Francescutti, hoy, con el escaso espacio disponible en los medios, un periodista tiene que pensarlo dos veces antes de cubrir un estudio de laboratorio. Francamente, no veo que estos trabajos justifiquen la mayoría de las veces una noticia destacada en un periódico general, o en un telediario, por más que trate del Alzheimer o de cualquier enfermedad de gran prevalencia, si no es debidamente contextualizada.

Cobo: Lo habéis resumido muy bien: es la paradoja entre el estudio inicial, que es sugestivo, y el final, que es confirmatorio de lo que ya sabíamos. Pero os devuelvo la pregunta: aunque el primer viaje de Cristóbal Colón fuera sólo tentativo, aunque tan sólo lanzase una posibilidad, aunque no estuviera confirmado su descubrimiento, ¿no os hubiera gustado a todos dar la noticia de que Colón regresó de las Indias?

Alonso: Eso no fue un hallazgo con ratones: fue un hallazgo espectacular, aunque Colón no entendiera bien a dónde había llegado. No fue tan preliminar como dices.

Cobo: Coged lo positivo de la analogía con Colón: cuanto más inicial es un descubrimiento, más peligroso resulta dar la noticia. Lanzaos, pero sed muy prudentes y decid enseguida que se trata de un estudio con animales, y que sólo dentro de bastantes años podría tener algún impacto en la salud humana.

Alonso: Pero al hacerlo estás introduciendo ruido en la población con mensajes que se confunden constantemente. No creo que sea el objetivo de la prensa general desinformar a la población con mensajes contradictorios, por muy atractivos que sean.

Público: A la gente le interesa la noticia que le ofrecemos en la medida en que le afectará en el futuro. Por eso la pregunta habitual al científico que presenta un estudio in vitro es «¿cuándo prevé que tendrá aplicación?». Y te suelen contestar «no te lo puedo decir».

Casino: Quienes hacen investigación básica en el ámbito de la salud son conscientes de que tienen menos posibilidades de obtener repercusión mediática. Y cuando el aparato de comunicación de sus centros envía a los medios un comunicado de prensa sobre su trabajo, los investigadores ya han entrado en la rueda mediática y muy a menudo acaban saliendo en los medios. Pero es preciso que digan a los periodistas «cuidado, esto es una investigación de laboratorio y no estamos hablando de nada que tenga que ver, por ahora, con la salud humana».

Cobo: Hace unos años se produjo una novedad magnífica. Un metaanálisis realizado por la Colaboración Cochrane concluyó que no existían evidencias para aconsejar el reposo en caso de lumbalgia; por el contrario, era aconsejable una actividad moderada. Es de agradecer que alguien hiciera la revisión y que hoy, gracias a ella, la lumbalgia dure menos. ¿Por qué la prensa no se hizo eco de una noticia tan importante? Sospecho que porque preferimos las noticias positivas. Si el metaanálisis, en lugar de decir que no hay que aconsejar el reposo para la lumbalgia, hubiera dicho que el ejercicio moderado puede mejorar la lumbalgia, hubiera ofrecido un resultado positivo y quizás habría tenido más impacto mediático. Por eso, lo que apuntó Valentín Fuster esta mañana me parece muy importante: debe-

mos ser positivos, y tanto más en el entorno en que nos movemos ahora.

Público: Para mí la diferencia entre informar o no de un hallazgo en ciencia básica depende mucho de la revista en que se publique. ¿Tengo que fiarme de la publicación? Porque yo me guío por la publicación: si un artículo sale en *Science* o *Nature* va a misa.

Casino: Si los artículos vinieran con una clasificación de confianza alta, media, baja y muy baja, las cosas serían más sencillas.

Alonso: Esa confianza está implícita en que sean publicados en *Nature* o *Science*.

Casino: Son dos jerarquías diferentes: la de la revista, si es de mayor o menor impacto, y la jerarquía de confianza según la calidad del estudio. Y luego hay una tercera, que a menudo es la que más importa para decidir si se informa o no: el interés periodístico.

Alonso: Mi consejo a un periodista es que si tienes un titular que te quema en las manos, prudencia, porque un gran descubrimiento médico ocurre en contadas ocasiones. Las revistas científicas también viven de la publicidad y necesitan titulares potentes. A veces las cinco grandes de medicina, *The Lancet, New England Journal of Medicine, British Medical Journal, JAMA* y los *Annals*, pecan de lo mismo que algunos periodistas y publican artículos fantásticos sin examinarlos con el rigor necesario. Se han publicado estudios interrumpidos antes de tiempo, sin criterios explicitados a priori para evaluar los resultados, con pocos eventos y pocos pacientes, en revistas de alto impacto, que muchas veces han acabado contradiciéndose, incluida *Nature*. Como periodista tienes que tener criterio propio, no puedes fiarte únicamente de las revistas.

García Molina: A propósito de este punto, Casino señaló un estudio epidemiológico defectuoso publicado en *Nature*, que no tiene gran tradición en esa disciplina.

Casino: Si se ha hecho un estudio de salud muy importante lo normal es que no se publique en *Nature*, sino en alguna de las cinco grandes revistas de medicina. Casi con seguridad, estas revistas habrían rechazado el estudio del que yo hablaba.

Cobo: Cuando tenéis que coger la evidencia existente y determinar hasta qué punto creer los resultados de un trabajo, os puede resultar muy útil la declaración Consort de ensayos clínicos y estudios observaciones. Las grandes revistas médicas piden a los autores de los artículos que digan en qué página o línea se está satisfaciendo cada uno de los ítems de esa declaración. Es una posibilidad que tenéis. Con respecto a lo dicho sobre *Nature*, totalmente de acuerdo; esta revista no sigue esas líneas de comprobación ni esas declaraciones, es de otro estilo y toca otros temas. Otra cosa: atentos al cambio en los estilos de publicación. Existe una iniciativa que traslada el coste de publicación al investigador, ya que él paga por publicar y el lector tiene libre acceso a los artículos. El modelo Plos, por ejemplo. Otro cambio llamativo concierne a las críticas de los revisores y las respuestas de los autores, que ahora se publican en algunas webs. Esto permite a los periodistas conocer las críticas que los revisores han hecho al estudio.

Francescutti: Sin duda, estas enseñanzas presentan el mayor interés, aunque resultan difíciles de aplicar incluso por periodistas especializados. Pero el problema mayor es que ya casi no quedan periodistas especializados en ciencia en las redacciones; en los medios sólo hay "todoterrenos", periodistas generalistas con menor capacidad de discriminación en estos temas. En contrapartida, prolifera la comunicación institucional de la ciencia, palpable en los *press releases* de los centros, instituciones y universidades que bombardean a los escasos medios que quedan. Y si llevar a la práctica esas enseñanzas en las redacciones parece difícil, no digamos ya aplicarlas en los gabinetes: veo muy difícil que sus miembros vayan a enmendar la plana a los científicos de sus centros y a sus superiores, ambos empeñados en dar la mayor repercusión a su investigación.

Casino: Sí, todo esto es un poco utópico… Así es como deberíamos funcionar los periodistas, en el supuesto de que existiera un escenario adecuado para conducirnos mejor.

Público: ¿Cuántos de los que estamos aquí hacemos periodismo?

Casino: Probablemente muy pocos; la mayoría se dedica a la comunicación.

Público: Quería preguntaros sobre el sesgo que supone la financiación de la mayoría de los estudios médicos por parte de la industria farmacéutica.

Alonso: Hay muchos trabajos que comparan resultados de estudios financiados por la industria con otros no financiados por ella, y detectan un sesgo conocido: en términos estadísticos, los estudios sobre fármacos financiados por la industria son más positivos que los demás. Mi consejo: si un artículo está financiado por laboratorios, examinadlo con más atención, e igual manteneos atentos al sesgo intelectual del investigador, que quiere dar la vuelta a sus datos para que sean noticiables y tengan más impacto.

Cobo: Soy de los pocos estadísticos que pueden presumir de haber participado en el desarrollo de dos productos que han llegado al mercado, algo muy difícil para un producto farmacéutico. El dueño de la compañía no tenía ningún interés en publicar los resultados de los ensayos clínicos de sus fármacos. En aquella época no había ningún apremio por publicarlos, pues lo que interesaba era que las agencias reguladoras los aprobasen. Después de esa aprobación ya tiene interés publicar los resultados. Si volvemos al ejemplo de los hipotensores, recordemos que salen al mercado tras demostrar que bajan la tensión. Más tarde, la FDA dice al fabricante «demuéstreme que su hipotensor también reduce el ictus». «De acuerdo –le responde el laboratorio– pero déjeme ponerlo en el mercado», y la FDA le da la autorización. De esta manera se inician los estudios de fase IV, ensayos que se realizan después de la comercialización en los que se estudian muchos aspectos de un fármaco que ya está en el mercado y que el paciente tiene que pagarse, con lo que la investigación le sale prácticamente gratis al laboratorio. Para concluir esta sesión, deciros que si un estudio resulta crucial para conseguir que el fármaco llegue al mercado, tomáoslo en serio; en cambio, si se trata de un estudio de un medicamento que ya está en el mercado o de uso compasivo, miradlo con otros ojos.

Un riesgo que crece del 1% al 2% puede significar un aumento del 1% o del 100%, según se mire

González / Ventura

Talleres
Análisis de *papers,* comunicados de prensa y artículos periodísticos

Esperanza García Molina

En la jornada de bioestadística para periodistas, después de los diálogos se llevaron a cabo talleres prácticos para analizar informaciones científicas con contenido estadístico, valorar su difusión en los medios de comunicación e identificar los problemas que estos casos podrían plantear a los periodistas en el ejercicio de su profesión. Para trabajar sobre estas cuestiones, los participantes en los talleres recibieron un *dossier* con varios ejemplos, cada uno de ellos compuesto por un artículo científico o *paper,* el comunicado de prensa redactado por la institución científica responsable de la investigación y artículos periodísticos reales escritos sobre el estudio. Además, en el *dossier* se planteaban una serie de cuestiones que los participantes debían resolver en cada ejercicio. Las preguntas eran las siguientes: ¿qué tipo de estudio es?, ¿cuál es el *outcome* o criterio de valoración?, ¿es adecuado el diseño del estudio?, ¿qué resultados ofrece?, ¿cuál es su mensaje principal?, ¿cuál es su nivel de confianza?, ¿es consecuente el *press release*?, ¿merece la pena informar?, ¿qué cautela habría que tener al informar? y ¿es correcto el titular periodístico?

Los participantes se organizaron en grupos de tres o cuatro personas con un representante para responder a las cuestiones que se planteaban en los talleres. Erik Cobo y Pablo Alonso fueron los dos expertos que dirigieron los talleres, junto al periodista Gonzalo Casino.

En este capítulo sobre los talleres prescindiremos de publicar íntegramente los textos que se utilizaron para no incurrir en violaciones de derechos de reproducción de los artículos, tanto los periodísticos como los científicos. No obstante,

facilitamos las referencias bibliográficas suficientes para encontrarlos. En el resumen de estos talleres se incluyen párrafos textuales, con las referencias a sus publicaciones, para poder seguir los ejemplos.

Aunque el *dossier* incluía más ejemplos para trabajar, durante los talleres sólo se escogieron tres. Por falta de tiempo, únicamente en el primero se completaron las 10 cuestiones planteadas; en los otros dos, nos limitamos a analizar en detalle el artículo científico, pero no su difusión en los medios.

Primer ejemplo: relación entre la miopía y la exposición a la luz en niños menores de 2 años

El primer ejemplo analizado en los talleres se basó en un artículo publicado en *Nature* en el año 1999 sobre miopía y medio ambiente, con el título *Myopia and ambient light at night.*[1]

Gonzalo Casino sugirió hacer primero una lectura del artículo en diagonal, pensando en una situación real de trabajo en la redacción, con pocas horas para valorar si hacerse eco o no de ese artículo y, en caso afirmativo, cómo hacerlo. Más tarde se respondieron las 10 preguntas para valorar el paso del *paper* a la noticia.

El artículo trataba sobre un estudio en padres de niños de 2 a 16 años de edad, pacientes de una clínica oftalmológica pediátrica. Los padres completaron un cuestionario acerca del nivel de exposición a la luz que habían tenido sus hijos antes de cumplir 2 años. Concluía que la prevalencia de la miopía en la infancia tenía una fuerte asociación estadística con la exposición a la luz durante el sueño en los primeros 2 años de vida, y que esta relación era dependiente de la dosis.

La primera cuestión planteada en el taller consistía en delimitar de qué tipo de estudio se trataba. Los participantes estaban de acuerdo en que no era un estudio de intervención, sino observacional. ¿Pero de qué tipo? ¿Cohortes, series de casos, o casos y controles? Se recordó que el criterio básico de un estudio de cohortes es que la muestra esté seleccionada por la exposición. Es un criterio clave que ayuda a distinguir el tipo de estudio. En este caso, la muestra se seleccionó tomando en cuenta a los pacientes que habían acudido a la consulta de un oftalmólogo y se les pidió que cumplimentaran una encuesta, en la cual se reflejaban hechos que habían sucedido con anterioridad. Erik Cobo aclaró que seleccionar a los participantes por la exposición no significa seleccionar sólo a los expuestos, sino escoger un grupo de individuos y medir en ese momento su grado de exposición.

Está claro, por tanto, que no se trata de un estudio de casos y controles. No se seleccionan los casos y después se buscan los controles, sino que se eligen pacientes, y de ellos unos tienen miopía y otros no. Más tarde se pregunta si de pequeños han estado expuestos a la luz durante el sueño. Como no han sido seleccionados en función de su exposición ni de su resultado (miope/sano), es una serie de casos. Las series de casos lo son independientemente de su tamaño: pueden ser de dos o tres casos, o de muchos. Aunque las dos variables se recogen en el mismo momento del tiempo, una de ellas hace referencia al pasado, por lo que más investigadores lo clasificarían como longitudinal retrospectivo que como transversal.

¿Cuál es el criterio de valoración en esta serie de casos o qué se está midiendo? La relación entre la miopía y la exposición a la luz. ¿Y qué se encuentra? Primero, que hay una asociación fuerte de la miopía con la exposición a la luz durante el sueño. Además, que es dependiente de la dosis, una de las cosas que pueden aumentar la credibilidad del estudio: cuanto mayor es la exposición a la luz, mayor es el riesgo de desarrollar miopía.

¿Es adecuado el diseño? Es un diseño poco fiable. En la pirámide que jerarquiza la calidad de los tipos de diseño está cerca de la base, y no permite establecer causalidad. Su nivel de confianza es muy bajo. Además, el estudio se ha hecho con una muestra de participantes obtenida de una clínica terciaria, que es una muestra difícilmente extrapolable a la población general.

¿Qué se echa de menos en la metodología? Que el estudio no se ajusta por la herencia familiar de miopía, pues no se estudia en cada caso si los padres son miopes o no. Además, el estudio se basa en lo que han respondido los

padres en un cuestionario sobre la exposición de sus hijos a luz hace muchos años; padres que, además, han llevado a sus hijos a un oftalmólogo. Una recogida de datos retrospectiva con un intervalo temporal en algunos casos de hasta 16 años es poco fiable. Gonzalo Casino señala que en este estudio interviene el sesgo de memoria. Una de las asistentes a la jornada añade que se está preguntando a padres cuyos hijos tienen problemas de visión si les dejaban la luz encendida de pequeños, y es fácil que cambien sus recuerdos en busca de justificaciones para la miopía de sus hijos.

En este punto, Erik Cobo interviene para señalar que el sesgo de memoria puede ser inocente, en el sentido de añadir solamente más variabilidad y ruido, o puede no ser inocente, «porque es posible que los padres de los niños que han desarrollado miopía recuerden detalles de ese tipo con mayor facilidad, puesto que se fijaron en ellos». La clave para hacer un estudio más fiable, añade Cobo, reside en que la recogida de las dos piezas de información sea lo más independiente posible, que no pueda haber una conexión entre ellas. Esto se consigue mediante enmascaramiento. En un estudio de cohortes prospectivo se recoge la información de las dos partes de forma enmascarada. Cuando se recogen datos sobre la exposición nadie sabe lo que pasará más tarde, y quienes toman los datos de la evolución no tienen delante los datos de exposición inicial.

¿Por qué hay relación entre miopía y dosis? Alonso comenta que puede ser causal o, por el contrario, deberse a que aquellos padres más miopes expusieran a sus hijos a mayores intensidades de luz porque ellos mismos la necesitaban. Es otro factor de confusión. ¿Por qué se publicó este estudio en *Nature*, una de las revistas científicas de mayor índice de impacto? Alonso responde que la revisión por pares de una revista no es un salvoconducto que garantice la fiabilidad de un estudio.

Entre el público, la periodista Ainhoa Iriberri insiste en que el hecho de que la muestra se haya recogido en una clínica resta validez al estudio. Cobo aclara que de esta manera se está reduciendo la ventana experimental, es decir, la perspectiva que se tiene en los datos. El espectro es más reducido y resulta más difícil encontrar relaciones, pero si se encuentran, la siguiente pregunta debe ser si la relación encontrada es aplicable a una ventana más amplia. Por otra parte, el análisis estadístico y el valor de p podrían ser correctos, ya que no se distingue entre relación y relación casual.

A continuación se analiza la nota de prensa que se preparó para difundir el estudio. En la nota no se informa de que los casos han sido seleccionados en una clínica. Un participante en el taller cree preocupante que la nota, por un lado, diga que no puede establecerse una relación de causa-efecto entre la exposición a la luz y la miopía, y sin embargo que por otro lado sugiera que no es aconsejable exponer a los niños a la luz mientras duermen. Cobo responde que si la nota dijera «los investigadores interpretan que los niños deben dormir con poca luz» sería impecable, porque es cierto que lo han interpretado así. Alonso añade que el titular de la nota de prensa es correcto: *Near sightedness in children linked to light exposure during sleep before age 2*. *Linked,* «relacionado» en inglés, es un término muy suave, pero no debería aconsejar un cambio de actitudes ante un estudio con tantos problemas de fiabilidad.

Seguidamente se pasó a analizar los artículos periodísticos. El periódico español *El País* dedicó el 17 de mayo de 1999 un reportaje a este estudio, con el título *Los bebés no deben dormir con luz,* un titular que, a la vista del análisis anterior, parece desafortunado. Alonso señala que la fuente independiente escogida por la autora del reportaje para contrastar la fiabilidad de los resultados, Eduard Estivill, es un médico especializado en temas de sueño, pero no oftalmológicos. En este punto, Casino incide en la importancia de consultar fuentes competentes en estadística. En este caso, haría falta un experto que supiera indicar al periodista el principal punto flaco del estudio: que no se ha controlado la herencia.

Casino llama la atención sobre el artículo publicado en el periódico británico *The Guardian*, con el título *Babies left in the dark see way to a brighter future* (13 de mayo de 1999), que contiene errores importantes de interpretación. «Está escrito por

Tim Radford, considerado uno de los mejores periodistas científicos de los últimos años; una prueba de que todos podemos meter la pata». Para Casino, lo más interesante de los errores que se cometieron con esta noticia es observar cómo a la prensa se le fue de las manos la información sobre un estudio que presentaba muchas debilidades. Alonso añade que hay que tener especial cuidado con noticias sobre mensajes muy novedosos con diseños observacionales pobres.

El artículo de *The Guardian* no contiene fuentes independientes del estudio, sólo cita a los propios autores. Es más, se atribuye declaraciones propias del autor que no son tales, porque son copia literal de la nota de prensa. Casino continúa poniendo como ejemplo de buena cobertura periodística el tratamiento que da el periódico estadounidense *The New York Times* a este artículo científico, al cual dedicó una noticia el 13 de mayo de 1999, a la vista del interés periodístico del tema, pero en el penúltimo párrafo indicaba que el estudio era prematuro, incompleto y que no había tenido en cuenta un factor obvio como la herencia. Es decir, se hace eco del estudio, pero lo analiza con fuentes independientes, especialistas en oftalmología, que supieron leer el artículo con perspectiva y ofrecer una visión crítica al periodista. En este sentido, Alonso opina que el periodista puede obtener una ayuda más valiosa del especialista al que consulta si, en lugar de darle a leer el *paper* sin más, le señala los puntos donde sospecha que puede haber problemas.

¿Los periodistas deberían haberse hecho eco del estudio o no? Casino opina que es cuestión de perspectivas, pero lo que hizo el *The New York Times* es perfecto. Además, el artículo se publicó sin firmar, lo cual es una lección de periodismo. Un participante reflexiona que, si medios como *El País* y *The Guardian*, con suficientes recursos, metieron la pata, ¿qué no pasará en medios menores o sin especialistas en ciencia en las redacciones? Para Casino, «lo fundamental es recordar que, en este caso, lo que falta es un buen conocimiento del contexto del problema, que no sabemos todavía cuáles son las razones que provocan la miopía». Una manera correcta de enfocar este estudio, según el periodista, es utilizarlo como percha para hablar de la miopía, recordar que no se conocen los factores que la provocan y plantear que este artículo ofrece una nueva hipótesis aún sin confirmar.

A favor de *El País* se recuerda que, más tarde, publicó un segundo reportaje sobre una revisión posterior en *Nature*, diciendo que no se habían podido demostrar las conclusiones del primer estudio y señalando sus limitaciones.

Segundo ejemplo: las pelirrojas sufren más miedo al ir al dentista

El artículo de este segundo caso se publicó en diciembre de 2012 en la revista *Journal of Endodontics*, con el título *Anesthetic efficacy of the inferior alveolar nerve block in red-haired women.*[2] El objetivo del estudio, que en su resumen se define como aleatorizado, era medir la posible relación entre la presencia del alelo de un gen (que tienen las personas pelirrojas) y la eficacia de la anestesia para el nervio alveolar inferior. Participaron en el estudio 62 mujeres pelirrojas y otras 62 con el pelo oscuro. Se les hizo responder un cuestionario para medir la ansiedad. Después se les administró la anestesia dental y se midió su efecto. No se encontraron diferencias en el efecto de la anestesia entre mujeres pelirrojas y morenas. Sin embargo, las pelirrojas manifestaron mayores niveles de ansiedad.

Este ejercicio se plantea, en primer lugar, con el objetivo de que los participantes en el taller sean capaces de averiguar el nivel de evidencia del estudio. Cobo recuerda la pirámide de la evidencia para analizar qué tipo de estudio es, y avisa de que tiene trampa. ¿Qué tipo de estudio es? Se trata de un estudio experimental. Es fácil reconocerlo porque en el texto aparece la palabra «aleatorizado».

¿Cuál es el *outcome*? La medida entre 0 y 100 del nivel de ansiedad de las participantes en el estudio mediante una escala de ansiedad. ¿Es adecuado el diseño? En principio, si es aleatorizado, parece que sí. ¿Qué resultados ofrece? Que las pelirrojas tienen más ansiedad al ir al dentista. ¿Pero en qué se sustenta esta afirmación? Realmente sólo en los resultados del test previo al experimento. En el estudio intentan es-

tablecer una relación entre la presencia de un gen concreto en las personas pelirrojas y una menor efectividad de la anestesia, pero los resultados de esta parte del ensayo indican que no existe dicha relación. La anestesia les hace el mismo efecto que a las morenas. Por lo tanto, el único resultado que hallan es la respuesta ansiosa que revela el test. Al leer el artículo se empieza a notar que hay algo raro.

Las conclusiones del artículo indican que el pelo pelirrojo y la presencia de este gen están ligados de manera significativa a un mayor nivel de ansiedad. Se comparan los resultados de una muestra en la cual la mitad de las mujeres son pelirrojas y la otra mitad no lo son; hasta ahí todo parece correcto, está equilibrado. Prosigue Cobo recordando qué significa la causalidad entre dos variables A y B: que al cambiar la variable A se consigue una modificación en la variable B. Pero, si no se puede variar A, ¿para qué preguntarse si la relación con B es causal? La pregunta clave en este caso es: ¿qué se ha aleatorizado en este estudio? No puede decidirse si una mujer será pelirroja o no, si tendrá o no un alelo.

En realidad, los investigadores están haciendo un estudio de asociación, e introduciendo aleatorización en una variable secundaria, de adorno según Cobo; esta variable secundaria es el lugar donde ponen la inyección de anestesia, en la mandíbula derecha o en la izquierda. Se confunde a los periodistas y a los editores de la revista haciendo creer que es un estudio aleatorizado, del máximo nivel de evidencia, cuando las variables de interés son observadas: el pelo pelirrojo, la presencia de un alelo y el nivel de ansiedad.

Cobo explica que normalmente pueden aleatorizarse pocas variables dentro del mismo estudio. Los estudios médicos aleatorizan una, que es la variable sobre la cual se hace la intervención. Pero el pelo pelirrojo es imposible de aleatorizar, porque no puede intervenirse en esta condición. Gonzalo Casino apunta otro detalle. El artículo se extiende al hablar de genética, cuando la publicación es una revista de endodoncia, en la que no escriben científicos sino endodoncistas. Es un estudio de baja calidad y nivel de evidencia, que tuvo mucho impacto en los medios por ser llamativo para el público y porque la nota de prensa se publicó en Eurekalert. Después de estas conclusiones, no se siguió analizando el caso y se pasó al siguiente ejemplo.

Tercer ejemplo: ¿son los efectos clínicos de la homeopatía efectos placebo?

El siguiente ejemplo es un artículo de revisión sobre la ineficacia de la homeopatía. Se publicó en *The Lancet* en 2005 con el título *Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy*.[3]

Los autores hicieron una revisión de ensayos clínicos con tratamientos homeopáticos y fármacos convencionales para estimar en qué medida podían estar afectados por sesgos. La búsqueda de estudios se realizó en 19 bases de datos, y la selección de ensayos en el Registro de Ensayos Clínicos Controlados Cochrane, asumiendo que aquellos que estuvieran realizados con doble ciego y aleatorizados eran de calidad superior. Al restringir el análisis a los estudios de alta calidad, se obtuvo una *odds ratio* de 0,88 con un intervalo de confianza de 0,65 a 1,19 para la homeopatía, y una *odds ratio* de 0,58 con un intervalo de confianza de 0,39 a 0,85 para la medicina convencional. La conclusión de este estudio es que hay una evidencia muy débil de los efectos específicos de la homeopatía y una fuerte evidencia de los efectos de los fármacos convencionales. Esto es compatible con la hipótesis de que los efectos de la homeopatía apenas se distinguen de los del placebo.

El análisis de este caso en el taller fue muy útil para que los periodistas entendiéramos el significado de los conceptos de *odds ratio* y de intervalo de confianza. Cobo explicó que, entre las medidas del riesgo, la más sencilla es el riesgo relativo. Se construye midiendo la proporción de personas que experimentan cierto suceso en un grupo de estudio, la misma proporción en un grupo de referencia, y después haciendo el cociente entre ambas.

Si, por ejemplo, se mide el efecto de un tratamiento frente al placebo, y vale 0,88, esto significa que el riesgo en los pacientes tratados dis-

minuye un 12% ([0,88 –1] x 100) respecto a los no tratados. Es decir, los pacientes sometidos a ese tratamiento tienen un 12% menos de riesgo de sufrir el suceso que los que toman placebo. En el caso que se analiza, el intervalo de confianza para las terapias homeopáticas es de 0,65 a 1,19. Comprende desde una reducción del riesgo del 35% ([0,65 –1] x 100) hasta un aumento del 19% ([1,19 –1] x 100), todo en términos relativos, no absolutos.

Continuamos con el ejercicio. ¿Qué tipo de estudio es? Es una revisión sistemática. Hay dos claves para reconocerlo. El estudio reúne lo que hay publicado sobre un tema, es una búsqueda exhaustiva. Además, evalúa la calidad de los estudios y para ello tiene en cuenta que los ensayos sean a doble ciego (para evitar el sesgo de realización) y que tengan la adecuada aleatorización (para eliminar el sesgo de selección). Se asume que esos estudios tienen una metodología de mayor calidad y se escogen porque los pacientes son dos grupos comparables, y ni pacientes ni médicos saben qué están tomando o prescribiendo.

¿Cuál es el criterio de valoración? La revisión compara ensayos clínicos de medicina convencional de múltiples intervenciones frente a placebo, con todos los ensayos que encuentra de intervención homeopática frente a placebo. De los homeopáticos hallaron más de 100 y buscaron otros tantos equivalentes de medicina convencional, que estudiaran las mismas enfermedades y tratamientos parecidos, aunque evaluaran cosas diferentes. Luego metaanalizaron con un análisis estadístico conjunto todos los resultados de los estudios que comparaban homeopatía frente a placebo, independientemente de lo que estuvieran comparando, y obtuvieron un estimador global. Hicieron lo mismo con todos los estudios de tratamiento convencional frente a placebo equivalentes a los de homeopatía, y así llegaron a tener dos estimadores.

Una de las conclusiones del artículo es que en ambos tipos de estudios, de homeopatía y de medicina convencional, los de baja calidad mostraron un beneficio importante, mayor que los de alta calidad. Puede que, por intereses comerciales, sólo se publiquen los casos positivos. Pero también puede haber un riesgo de sesgo porque los ensayos no estén bien aleatorizados o controlados.

No obstante, el análisis se hizo únicamente para los estudios de alta calidad. La *odds ratio* en los estudios de homeopatía, con un valor de 0,88, es no significativa, pues sólo representa una mejoría del 12% respecto al placebo, explica Alonso. El intervalo de confianza, de 0,65 a 1,19, indica que la precisión es baja. Muestra tanto importantes beneficios como perjuicios, por lo que estos estudios no son clínicamente relevantes. En este punto, Alonso aclara que cuando un intervalo de confianza incluye el 1, esto quiere decir que el tratamiento estudiado puede no tener ningún efecto positivo respecto al placebo.

Cobo apunta que esto no quiere decir que ese tratamiento sea perjudicial, sino que el efecto nulo es uno de los valores compatibles con los resultados observados para ese tratamiento; pero hay otros, unos positivos y otros negativos. «Aquí aplica aquello de que ausencia de pruebas no es prueba de ausencia», aclara Cobo. Alonso recalca que, para que pueda afirmarse que un tratamiento reduce el riesgo, el intervalo de confianza debe estar por debajo de 1. No obstante, aclara que, aun así, puede no ser clínicamente relevante por reducir muy poco el riesgo, resultar muy caro, tener efectos secundarios, etcétera. El umbral varía en función de los efectos adversos y de los inconvenientes.

Seguimos analizando el intervalo de confianza en estudios equivalentes de alta calidad en medicina convencional. Para estos casos, la *odds ratio* es 0,58, lo que significa que los tratamientos convencionales reducen el riesgo un 42% respecto al placebo. El intervalo de confianza (0,39 a 0,85) está, en el peor de los casos, por debajo de la unidad. La medicina convencional demuestra tener resultados positivos.

El intervalo de confianza cuantifica la magnitud de la incertidumbre, añade Cobo, pero no debe extraerse de él la idea de que el efecto es diferente en distintos pacientes. Una de las premisas es que el efecto es el mismo en una misma población. Asumiendo que el efecto es el mismo para todos los pacientes, el intervalo de confianza indica cuáles son los valores plausibles para este efecto común en una población.

Finalmente, se terminó este ejercicio con la conclusión de que es una revisión sistemática con un alto nivel de confianza.

Bibliografía

1. Quinn GE, Shin CH, Maguire MG, Stone RA. Myopia and ambient lighting at night. Nature. 1999;399:113-4.
2. Droll B, Drum M, Nusstein J, Reader A, Beck M. Anesthetic efficacy of the inferior alveolar nerve block in red-haired women. J Endod. 2012;38:1564-9.
3. Shang A, Huwiler-Müntener K, Nartey L, Jüni P, Dörig S, Sterne JA, et al. Are the clinical effects of homoeopathy placebo effects? Comparative study of placebo-controlled trials of homoeopathy and allopathy. Lancet. 2005;366:726-32.

Las condiciones no son causas útiles

Cobo / Ventura

problems with media coverage and how to do better

Steven Woloshin and Lisa M. Schwartz

The two basic ingredients of good decision making are facts and values. Facts refer to the available choices and the likely outcomes of their choices. Values refer to how much people care about the different outcomes and the associated tradeoffs (e.g., the potential benefits and harms, costs and inconveniences of their choices). People can only make good decision when they have the facts and some clarity about their values.

This simple model of decision making highlights a basic problem: without the facts, people cannot possibly make good decisions. They may make a lucky decision and have a good outcome, but not an informed decision consistent with their values.

When it comes to medical care, people see lots of messages. Unfortunately, many do not provide the facts. Consider this colon cancer screening advertisement from Sloan Kettering –a major cancer hospital in New York– which ran in the *New York Times* (fig. 1). This ad does not present facts, it presents fear. It says be afraid: you may feel healthy, but guess what, you may have colon cancer. It says you can never feel safe because even when you seem well you may really be sick. The ad is also highly exaggerated

Figure 1.

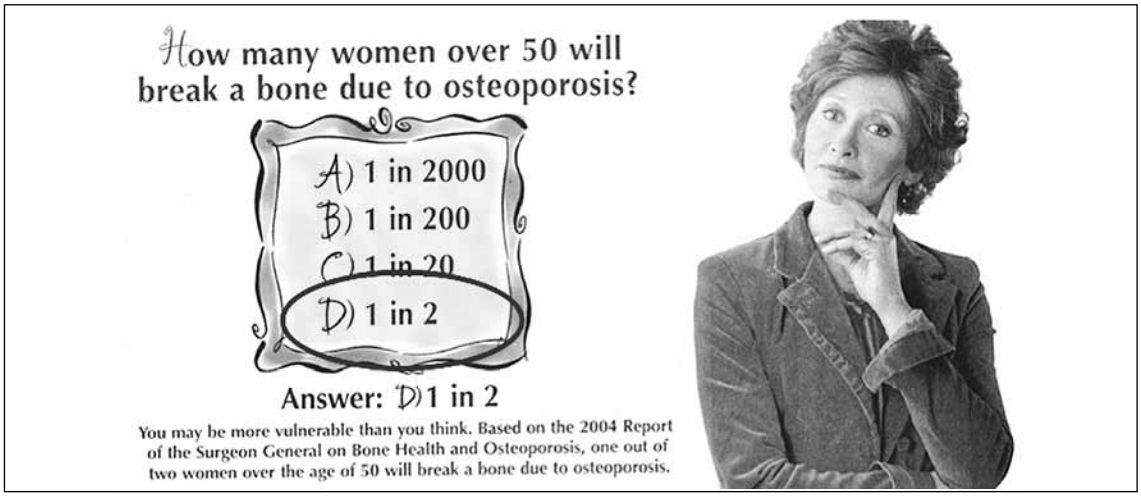


Figure 2.

since most 50 year olds who feel great and have a healthy appetite do not have –and will not get– colon cancer. For example, on average, a 50-year-old man has a 3 out of 1,000 chance of being diagnosed with colon cancer in the next 10 years and 1 out of 1,000 chance of dying from it.

Or consider this direct-to-consumer advertisement (fig. 2) promoting a drug for osteoporosis –thinning of bones (only the US and New Zealand permit direct to consumer advertising of prescription drugs– in the US, citizens are exposed to over $4 billion of these ads each year, ten times the FDA's budget for evaluating new drugs).

This ad looks like it presents facts –but it is an illusion of facts. The "1 in 2" number greatly exaggerates a woman's risk of fracture. The "1 in 2" number includes both fractures that hurt (and cause problems) and fractures that are small (which can only be seen on x-rays and never cause symptoms or problems). But most importantly, the vast majority of fractures from osteoporosis occur among women 75 and older –not among women 50 to 75. The message in the fine print reveals the true purpose of the ad: to make women feel vulnerable and afraid. The print under the numbers says: *"You may be more vulnerable than you think"* (Fig. 2).

That a drug company might exaggerate risk to sell a product is not so surprising. Seeing the same tactic from a disease awareness group is. The Light of Life Foundation (a disease awareness group founded by a thyroid cancer survivor) ran a series of ads to promote thyroid cancer screening (fig. 3).

The ad depicts Rachel, age 14, "the day before she was diagnosed with thyroid cancer".

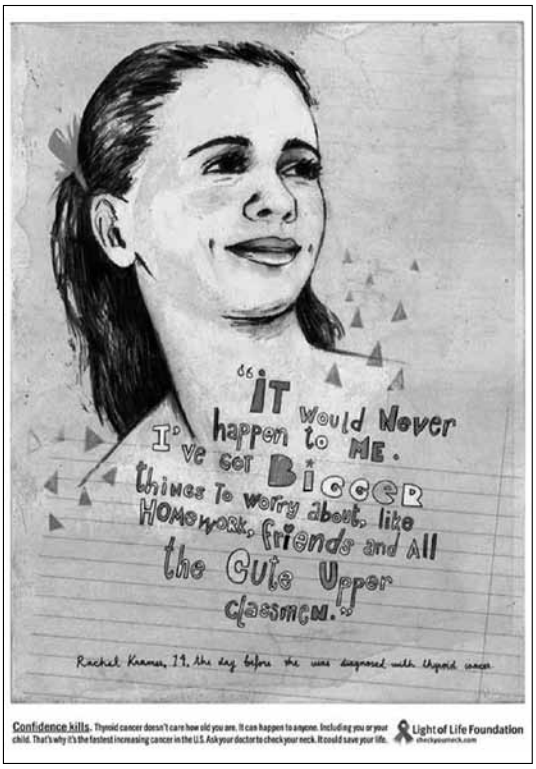


Figure 3.

Rachel says: *"It would never happen to me. I've got bigger things to worry about like homework, friends and all the cute upper classmen".* And the ad's bottom line reads *"Confidence kills: Thyroid cancer doesn't care how old you are. It can happen to anyone. Including you or your child".* We think this use of "facts" to generate fear in young people and their parents is actually cruel. A 15-year-old girl's chance of getting thyroid cancer in the next 10 years is less than 1 out of 1,000 and the chance of dying from it about 1 out a million. Because the disease is so rare, no professional medical organization recommends thyroid cancer screening for young girls.

The prior three messages share a pattern: using hype to generate extreme fear. But many messages go in the opposite direction. They use hype to generate extreme hope.

In December 2003, the cover of the magazine, U.S. News and World Report, declared "The end of heart disease" (which as of this writing in 2013 remains the biggest killer in the United States). Extreme hope also comes from leaders of our most esteemed organizations. During the U.S. National Institutes of Health budget hearings in 2005, a senator asked Dr. Von Eschenbach, the director of the National Cancer Institute at the time, *"What is going to happen by 2015 as you project it?"* The directors responded, *"No one who hears the words 'You have cancer', will suffer or die from the disease. We will prevent and eliminate the outcome"*.[1] Unfortunately, despite receiving their requested budget, the National Cancer Institute is, of course, nowhere near eliminating suffering or death from cancer.

The most effective "message" strategy is to use fear and hope together: exaggerate a risk to make people feel vulnerable and then exaggerate the benefit of what you have to offer (or sell) to reduce that risk. The advertisement (fig. 4) for abdominal aneurysm surgery from Mount Sinai (a major academic medical center in New York) illustrates the power of this strategy.

The ad makes two statements. One generates fear: an aneurysm is a death sentence. The other generates hope: Mount Sinai can offer a pardon. But both statements are highly exaggerated. Most aneurysms that are found are very

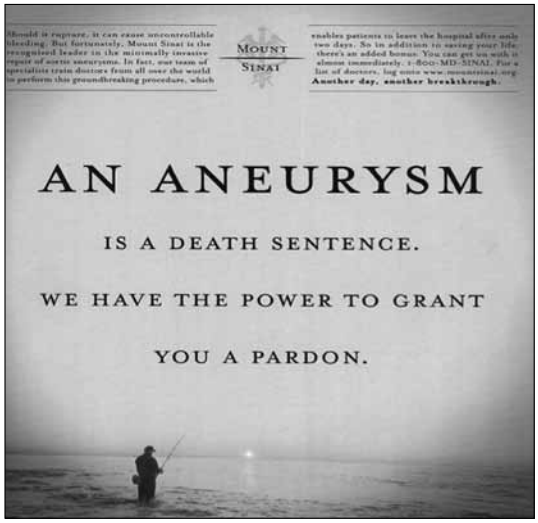


Figure 4.

small and will never grow large enough to cause problems –let alone death. And a few people will die from the surgery –not exactly a pardon.

The problem is that messages exaggerating disease risk and treatment benefit are everywhere. The problem is that these messages cause harm. They generate anxiety and undermine the public's sense of well-being and resilience. As a result, they may prompt too much exposure to medical care which may not help and can really hurt people. And repeated exposure to exaggeration may leave the public cynical: they may stop paying attention to health messages altogether.

It is easy to understand why there is so much exaggeration. Manufacturers (drug, technology) need to sell their products. Academic institutions need publicity to raise funds. Meeting organizers need to attract scientists, advertisers and sponsors. Researchers need to show results to advance their careers. Media outlets compete for stories, advertisers, and readers (or viewers). And journalists compete for the front page –or the most e-mailed story. This is a recipe for exaggeration because so many self-interests are served by being associated with research perceived to be new, big and important.

If the diagnosis is exaggeration, we think the prescription is healthy skepticism. We all need to push back and see through exaggeration to avoid being manipulated by messages that make

us too scared or too hopeful. This is especially important for journalists. The media's power to amplify and disseminate messages makes journalists a prime target for exaggeration.

Journalists' sources exaggerate

Exaggeration often begins with journalist's sources. Researchers do this when they suggest their findings apply to more people than they really do. Or when they are too certain about inherently weak science and fail to acknowledge study limitations. One way this plays out is through researcher quotes, a feature seen in almost all press releases. In a systematic review of press releases issued by academic medical centers, we judged one-quarter of researcher quotes as "exaggerated".[2] For example, in a press release titled "Scientists inhibit cancer gene. Potential therapy for up to 30% of human tumors", the lead investigator, said, "the implication is that a drug therapy could be developed to reduce tumors caused by Ras without significant side effects". The researcher greatly exaggerated the implications of this study since it only involved skin cancer in mice (no human testing for efficacy or safety had been done).

In the same systematic review, we documented other problems with the press releases –the most direct way that academic medical centers communicate with journalists.[2] Over one-third of the press releases failed to quantify the main result. When results were quantified, over half used formats known to exaggerate the magnitude of findings (for example, giving a relative change without providing the base rate). Despite the fact that all studies have limitations, few press releases mentioned them.

Many press releases promoted animal or lab research and specifically claimed that these studies were relevant to human health. Nearly all (98%) failed to caution about problems translating such research to humans. The need for caution was highlighted in a systematic review of "high profile" animal studies.[3] On average, it took 14 years to translate the animal research into human testing. And only one-third of animal studies translated into successful interventions in randomized trials of humans. Moving from animals to humans is a slow and uncertain process.

After finding similar problems in medical journal press releases, we interviewed press officers at major medical journals to better understand how releases are written. The interviews gave us insight into why press releases can be so problematic. None of the journals had data presentation standards for press releases. Nor did any require a statement about study limitations.[4]

Medical journals, academic medical centers and researchers –all important sources for journalists– contribute to the problems with health news. We now look at how these problems play out in what is actually reported in the media. We focus specifically on reporting in two high-risk zones for exaggeration: scientific meeting presentations and disease mongering.

Too much, too soon: media reporting on scientific meetings

Reporters routinely cover scientific meetings. These meetings, sponsored by large professional organizations, have two purposes: they are a forum for scientists to present work to colleagues and represent an engine for generating media coverage. In fact, the effort courting the media is often greater than the effort in vetting the sciences. In 2002 (when we conducted a study of media coverage of scientific meetings), the Society for Neuroscience received 15,000 abstracts for presentation and accepted 100% of them (the only criteria for acceptance was membership of one of the authors).[5] The only review conducted by the Neuroscience organization was to determine which abstracts would be promoted to the media.

Scientific meeting research is typically preliminary, may have limited relevance to human health, and has generally undergone limited –if any– peer review (i.e., reviewers typically have access only to the abstract –not the full manuscript). Nevertheless, research presented at scientific meetings is often big news.

To gauge the quality of these reports, we did a content analysis of news stories after five major scientific meetings (World AIDS, American Society of Clinical Oncology, Radiological Society

of North America, American Heart Association and the Society for Neuroscience).[6] We identified 174 newspaper stories (34 on the front page) and 13 national radio or television stories in the 2 months after the meeting. These stories appeared in 50 major news outlets, including eight of the top ten circulation U.S. newspapers. The bottom line was that there was lots of room for improvement. Basic study facts were often missing: one-third of the news stories failed to report the study size; about half did not state the study design and 40% did not quantify the main result. Cautions about studies with obvious limitations were also missing: all failed to caution about assuming the results of animal or cell research apply to human health, 69% failed to caution that in uncontrolled studies you cannot know if the intervention caused the finding and over half failed to caution about the instability of results from small (less than 30 patient) studies.

Remarkably, the most important caution about scientific meeting research –that it is preliminary, unpublished, not the final study results– was missing from all but one news story. This caution matters because preliminary work does not always pan out. Result change and fatal flaws emerge. These problems are reflected in the publication fate of scientific meeting research which garnered high profile media coverage. While half of this research is published in high-impact medical journals in the next 3 year, one quarter is published in low-impact journals and another quarter is never published.[5] This finding was the same for meeting research that covered on the front page of the newspaper. Dr. Richard Klausner, the former director of the National Cancer Institute, captured this phenomenon even better than the numbers: *"I'm pretty well plugged in to what's going on in research," he remarked. "I hear on the news "Major breakthrough in cancer!" And I think, Gee, I haven't heard anything major recently. Then I listen to the broadcast and realize that I've never heard of this breakthrough. And then I never hear of it again."*

The media and disease mongering

It is hard to avoid becoming sick. If you sleep too little at night, you have insomnia. But if you sleep too much during the day, you have excessive daytime sleepiness syndrome. If you have trouble paying attention, you have attention deficit disorder. But if you pay too much attention, you have obsessive-compulsive disorder. And if you have any blood sugar, blood pressure or even any bones, you may have pre-diabetes, pre-hypertension or osteopenia.

Diagnosis is expanding. We are turning ordinary experiences (such as transient sleep problems, sadness) into disease and turning risk factors into diseases themselves (such as high cholesterol, a risk factor for the heart attack is now a diagnosis itself with its own ICD9 code, etc.). And in either case, lowering the cutoff necessary for the diagnosis can expand an existing disease.[7] The late 1990's the threshold for being "overweight" was changed from a body mass index ≥25 kg/m^2 instead of ≥27 kg/m^2.[8]

Expanding diagnosis reflects a fundamental problem in medicine: how do we define sickness? Most medical phenomena exist on a spectrum. At one end, people are overtly sick. At the other end, people are perfectly well. A narrow definition of sickness –drawing the line closest to "overtly sick"– labels the few people with the diagnosis. The advantage is that the definition focuses on the sickest people –those who stand to benefit the most from treatment. The disadvantage is that we miss some people who might benefit. Ideally, we would draw the line based on the benefits and harms to patients. In reality, many forces –drug companies, device manufacturers and doctors– are pushing the line to create broader and broader definitions of sickness. Whether or not it helps patients, broadening disease definitions serves other interests.

Disease mongering is the effort to convince people that they are "sick" and need a medical treatment for this sickness. This means creating very broad definitions of disease and conducting disease awareness campaigns to raise undue concern about the prevalence and severity of "disease" to capture the biggest market. Disease mongering implies that this is being done for reasons other than the patient's interest.

The problem is that disease mongering can really make people sick. The anxiety, sick role from

the diagnosis and side effects from treatment can be worse than the disease. The primary culprits are drug companies who conduct disease promotion campaigns, run direct-to-consumer drug ads, fund disease advocacy groups, subsidize physician education (CME, etc.) and pay for clinical trials. But the facilitator is the news media. They are a highly visible source of health information for consumers (and physicians and policymakers). Because the news is more credible than advertisements, it is met with less skepticism. To avoid being co-opted into the process, journalists need to know how to recognize the signs and symptoms of disease mongering.

Case study: a drug in search of a new use and how the media helped

Years ago, GlaxoSmithKline developed a drug called *Requip*. It was a drug for Parkinson's disease, but not a very successful –a third line drug– and it was going off-patent. There were some reports that *Requip* could be used for an obscure movement disorder called Ekbom's syndrome (now known as restless legs syndrome). We are going to show how GlaxoSmithKline extended Requip's patent protection by turning this obscure disorder into –according to their direct to consumer drug ads– a "recognized medical condition. One shared by nearly 1 in 10 U.S. adults".[9]

The story actually began in the late 90's, when the International Restless Legs Foundation (a group of mostly industry funded scientists) created the definition of restless legs syndrome.[10] To have the diagnosis, a patient must have each of the four standard criteria:

1) An urge to move the legs due to an unpleasant feeling in the legs.

2) Onset or worsening of symptoms when at rest or not moving around frequently.

3) Partial or complete relief by movement (e.g., walking) for as long as the movement continues.

4) Symptoms which occur primarily at night and which can interfere with sleep or rest.

Treatment is reserved for those with moderate-severe symptoms judged by frequency.

In 2003, GlaxoSmithKline sought FDA approval of *Requip* for restless legs. FDA review generally takes about one year. Toward the end of this period, GlaxoSmithKline began launching a press campaign, beginning with press releases from presentations at the American Academy of Neurology meeting and a press releases from a company funded (and unpublished survey): "New survey reveals common yet under recognized disorder –restless legs– is keeping America awake at night". But FDA refused to approve the drug because the submitted studies were too short (12 weeks) raising questions about long term safety. The drug was finally approved in 2005 after Glaxo submitted a "long-term" study of 36 weeks. With approval, the drug company sought to "push restless legs syndrome into the consciousness of doctors and consumers alike" and began a 27 million U.S. dollar direct-to-consumer advertising campaign. Within a year, drug sales increased from $97 to $146 million.

To explore the role of the news media, we looked at coverage of *Requip* during the campaign.[9] We identified and rated the 33 news stories that appeared in major newspapers. Two-thirds of news stories simply repeated the "nearly 1 in 10 U.S. adults" prevalence estimates asserted in the drug company ads (a more critical reading of the prevalence studies suggests that <3% might need treatment, although even this number is likely to be an overestimate). Three-quarters of news stories discussed the extreme physical and emotional aspects (typically with a patient anecdote), yet none presented any anecdote of mild disease. Forty-five percent blamed doctors for being unaware of the diagnosis (e.g., "relatively few doctors know about restless legs. This is the most common disorder your doctor has never heard of") or suggested that patients were unaware they were sick. One quarter referred readers to checklists for self-diagnosis and for, for more information, to the "not for profit" Restless Legs Foundation. While the Foundation is "not for profit", its annual report discloses that, by far, the major funder is GlaxoSmithKline –the makers of *Requip*. Glaxo is the only gold medal

donor listed (defined as a minimum of a quarter of a million dollars –Pfizer, who had another restless legs drug in development was listed as a bronze medal donor). No news report mentioned entanglement between Glaxo (or Pfizer) and the Foundation. None of the news stories mentioned the possibility that there might be too much diagnosis.

Did the media accurately portray the benefits and harms of *Requip*? Not exactly. Among the 15 stories that mentioned Requip specifically, 45% only discussed the benefit of the drug with an anecdote and 33% used "miracle" language (for example, literally quoting a patient as saying "[*Requip*] has been a miracle drug for me"). Only one story quantified the benefit. The best estimate of the benefit of *Requip* comes from the 12-week randomized trial that was part of the basis of FDA approval. For the primary outcome (average improvement on the International Restless legs symptom score), the *Requip* group improved by 14 points vs. 10 point improvement in the placebo group, a net 4-point improvement on a 40-point scale. Is a change of this magnitude a "miracle"? Understanding what the 4-point change means is a challenge. Would patients notice it (power calculations in approval trials asserted 3 and 6 points as meaningful changes)? The study also looked at whether clinicians rated patients as "very much" or "much" improved: 73% of the *Requip* group improved compared to 57% of the placebo group –so only 16% of patients improved because of *Requip*.

Media reporting of the harms of *Requip* was also poor: only about one quarter mentioned any harm. But *Requip* has important side effects: nausea (40% vs. 8% placebo), dizziness (11% vs. 5%), somnolence (12% vs. 6%), and fatigue (8% vs. 4%). Increased chance of somnolence and fatigue undermine the rationale for the drug's use since much of the push for treating restless legs is how it is "keeping America awake at night". How useful is a treatment that improves restless legs symptoms for a minority of patients if it leaves almost as many feeling more tired and more fatigued. For some patients, in fact, the problem of tiredness was so severe that FDA required that *Requip* include a warning in the ad that "*Requip* has been associated with sedating effects, including somnolence and the possibility of falling asleep while engaged in activities of daily living".

In summary, the media did aid and abet disease mongering efforts. While restless legs syndrome is just one example, there is no reason to think other disease promotions would be covered any differently. Journalists can do better by being skeptical when new –or expanded– diseases are being promoted to them. More specifically, they can (and should) question prevalence estimates, present the full spectrum of disease, question the idea that more diagnosis is always better and quantify the benefits and harms of the new treatment. To help them do so, it is helpful to consult experts without financial or professional conflicts of interest (a list of "Industry-independent experts" is available at: http://www.healthnewsreview.org/toolkit/independent-experts/).

Conclusion

Problems with media coverage can have important consequences for the public. People may become too enthusiastic about new and marginally effective interventions and too certain about findings based on weak science. There are a number of ways for journalists help readers get the facts. One way is to report numbers. When journalists quantify the chance of disease and the benefits and harms of treatment, the public can appreciate the actual magnitude of the risks they face and decide whether the benefits of interventions outweigh the harms. Journalists should also routinely note important study limitations to help inoculate the public against believing that we know more than we do, and constrain unrealistic expectations. The tip sheets [see Apéndice, p. 75] we have developed provide cautions specific to common study limitations. Journalists with a healthy skepticism will promote a healthier public.

References

1. Goldberg K. Money would speed progress, NCI says, but backs off meeting 2015 goal by 2010. Cancer Letter. 2005; Vol 31. Washington, DC:1.

2. Woloshin S, Schwartz L, Casella S, Kennedy A, Larson R. Press releases by academic medical centers: not so academic? Ann Intern Med. 2009;150:613-8.
3. Hackam DG, Redelmeier DA. Translation of evidence from animals to humans. JAMA. 2006;296:1731-2.
4. Woloshin S, Schwartz L. Press releases: translating research into news. JAMA. 2002;287:2856-8.
5. Schwartz L, Woloshin S, Baczek L. Media coverage of scientific meetings: too much, too soon? JAMA. 2002;287:2859-63.
6. Woloshin S, Schwartz L. Media reporting on research presented at scientific meetings: more caution needed. Med J Aust. 2006;184:576-80.
7. Schwartz L, Woloshin S. Changing disease definitions: implications for disease prevalence. Analysis of the third National Health and Nutrition Examination survey, 1988-1994. Eff Clin Pract. 1999;2:76-85.
8. National Heart, Lung and Blood Institute. Clinical guidelines on the identification, evaluation and treatment of overweight and obesity in adults. 1998. Available from: http://www.ncbi.nlm.nih.gov/books/NBK2003/
9. Woloshin S, Schwartz L. Giving legs to restles legs: a case study of how the media helps make people sick. PLoS Med. 2006;3:452-5.
10. Allen R, Picchietti D, Hening W, Trenkwalder C, Walters A, Montplaisi J. Restless legs syndrome: diagnostic criteria, special considerations, and epidemiology. A report from the restless legs syndrome diagnosis and epidemiology workshop at the National Institutes of Health. Sleep Medicine. 2003;4:101-19.

Los modelos de previsión reducen la incertidumbre, no la anulan

Elmore / Ventura

Seeing through the hype: Garbage! When the news is not fit to print

Steven Woloshin and Lisa M. Schwartz

Sometimes the medical news may not be fit to print: the research is so preliminary or so inherently weak that reporting it would more likely mislead than inform the public. For example, stories about a new miracle diet, or cancer breakthrough –stories which often arise out of abstracts presented at scientific meetings that have not undergone peer review, or uncontrolled human studies. Nevertheless, journalists often feel pressure to report on such studies.

What should journalists do when the news is not fit to print? Ideally they would not report it. But if they have to report it, they should always include strong cautions to alert readers to questions about the validity, meaning or generalizability of the research. In this essay, we will use real examples to review cautions about research that may sound exciting but is very preliminary or inherently weak.

"Raloxifene may decrease the risk of endometrial cancer in post-menopausal women" Meeting abstract, American Society of Clinical Oncology 1998 meeting

When this abstract was presented at a plenary session at the 1998 ASCO meeting, several news outlets covered the story, including the *Wall Street Journal*.[1] The abstract reported on what seemed to be a major advance: a relatively new medication which behaved differently than others in its class which increased the risk of endometrial (uterine cancer).

The investigators presented results from a randomized trial of raloxifene, a selective estrogen receptor modulator, or placebo which included 7704 postmenopausal women (mean age 66.5 years) with osteoporosis (based on hip or

spine bone density at least 2 standard deviations below normal or a history of vertebral fractures). The study was "designed to test the hypothesis that women assigned to receive raloxifene will have a lower risk of fractures than women assigned to placebo". But the main result reported in the ASCO abstract was not about the primary outcome –osteoporosis– but about endometrial cancer. Of course, endometrial cancer is not the same as osteoporotic fractures.

While it is perfectly legitimate for a study to report on multiple outcomes, these outcomes need to be specified in advance. Otherwise, surprise findings –which may reflect chance alone may be over interpreted. That is, taken as strong evidence for a treatment effect rather than a hypothesis generating finding which needs confirmation in a subsequent trial. For example, if 20 random outcomes are assessed after a trial is completed, 1 will be statistically significant –have a p value less than 0.05 – just by chance. While endometrial cancer was a pre-specified study outcome, the researchers' hypothesis was the opposite: they were concerned it might increase cancer, a harm associated with raloxifene's rival, the drug tamoxifen. In fact, according to the trial protocol, assessing the drug's effect on endometrial cancer was designated as a safety issue, not as a potential benefit.[2]

Contrary to their hypothesis, they found that raloxifene *decreased* endometrial cancer: since this decrease in endometrial cancer was not a pre-specified hypothesis, the finding should be considered a "surprise" and interpreted cautiously.

How much did raloxifene decrease the risk of endometrial cancer? According to the abstract, "compared with the rate in the placebo group, the overall relative risk of endometrial cancer is 0.38 (p = 0.232). If two cases diagnosed within one month of randomization are excluded, the estimate of relative risk is 0.13 (p = 0.045)". While one may question the legitimacy of excluding data (particularly when doing so generates a desirable finding), the rationale for exclusion may be reasonable –the drug is unlikely to have had such a rapid effect, so the two cancers were probably present at the time of randomization. It would be reassuring to know that such exclusions were written into the study protocol, rather than excluded after the data were analyzed (we have not been able to ascertain which was the case in this trial). Another theoretical concern about this finding is whether the scrutiny was similar for both the raloxifene and placebo groups. If not, the investigators will have introduced bias. For example, if researchers looked harder for endometrial cancers in women in the raloxifene group compared to the placebo group (and excluded any found) this would bias the study in favor of the drug because preexisting cancers would still count against placebo. Fortunately, based on the protocol, the level of scrutiny was appeared to be the same in both groups in this trial.

Assuming the exclusions were legitimate, the fact that removing two cases had such a dramatic effect highlights a second concern: the results are very unstable. The magnitude of the risk reduction increased from 62% (relative risk = 0.38) to 87% (relative risk = 0.13) and the p-value changed from *not* statistically significant (p = 0.232) to statistically significant (p = 0.045). This instability reflects the preliminary nature of the report. The study was not over –only a fraction of the final data was collected. While the researchers did not report the number of cancers found in either the raloxifene or placebo groups, the number must have been very small. Waiting a little longer to accrue more data, a few more cases might flip the findings back.

In fact, waiting for more data did reverse the finding. When final study results were reported in *JAMA,*[3] raloxifene no longer had any effect on uterine cancer (Table 1).

Not only were the findings different, the same authors made a big change to their message. In the scientific meeting abstract, they wrote that the "Raloxifene *may decrease* the risk of endometrial cancer in post-menopausal women". In the *JAMA* article published 1 year later, they wrote "Raloxifene *did not increase* the risk of endometrial cancer". The *JAMA* message reflects their original hypothesis that endometrial cancer was a safety concern.

Here is how we would have rewritten the findings presented at the scientific meeting: *"There was a trend toward a lower rate of uterine cancer*

Table 1.

	Relative Risk (Raloxifene vs Placebo)	P-Value
1998 ASCO Meeting		
All data	0.38	0.232
Exclude 2 cases*	0.13	0.045
1999 JAMA	0.80	0.67

*Cases which ocurred within 1 month of randomization.

but it may be due to chance and it is too early in the study to say. This was not the main outcome being studied. We are not very confident in these results".

This example highlights the fundamental problem with early results –they turn out not to be true (as in this case) or they may change substantially. When they report on scientific meeting presentations, journalists should raise a red flag for their readers. Our suggestion for this cautionary note is "These preliminary findings may change because the study has not been independently vetted through peer review and all the data are not yet in". (Note: this caution and the ones that follow are summarized in the tip sheet: *How to highlight study cautions.* See Apéndice, p. 80.)

"Drug advances bring new hope to cancer battle –New treatments block 'switches' that turn cells malignant" *Wall Street Journal*

Other major U.S newspapers echoed the excitement of the *Wall Street Journal* headline: "Drug shrinks lung tumors" *Washington Post,* "Major step in cancer fight" *Houston Chronicle,* "Pill shows significant results in battling advanced lung cancer" *The Milwaukee Journal Sentinel*. Lung cancer is a terrible disease, one for which we do not have very effective treatments, so a real breakthrough would be wonderful news. Is this new drug really a breakthrough?

To understand whether this enthusiasm is warranted involves looking at the science behind the headlines.[4] The study followed 216 patients with advanced lung cancer who were all given the new treatment –*Iressa*. Unfortunately, there was no control group. Without a control group, it is extremely difficult to learn much about how well the treatment works. It is possible that equivalent patients who did not get the drug would have done worse (meaning *Iressa* helped). But it is also possible that the patients would have done no better or even worse (meaning *Iressa* caused harm). Here's the first red flag for journalists and readers –"Because everyone took Iressa, it is extremely hard to know if *Iressa* had anything to do with the outcome".

Even if this study were a randomized trial, a second fundamental problem exists: the primary outcome was a surrogate measure –tumor shrinkage (by half or more). The study found that 10% of the 216 patients had tumor shrinkage.[5] It is a big leap of faith to assume that tumor shrinkage means less suffering or death from lung cancer. There are three reasons why this is a leap of faith. Tumor shrinkage may be followed by period rapid growth. Or the tumor may shrink in an unimportant area that does not affect a person's health. Finally, spread in rest of body may be much more important than tumor shrinkage. So again, readers need a cautionary note: "This study measured tumor shrinkage –an x-ray finding that patients do not directly experience. Be cautious about acting on these findings since changes in these kinds of measures don't reliably translate into people feeling better or living longer".

Despite these two fundamental limitations, the study received a lot of enthusiastic press. This initial press coverage illustrates the beginning of an unfortunately common cycle. The cycle begins with great news –typically with breathless excitement about a new technology. But terrible news quickly follows –when side effects start to emerge as more people take the drug. In the case of *Iressa,* this happened with reports of drug-related

deaths in Japan where the drug was already approved for the treatment of advanced small cell lung cancer. Here is an excerpt from the *Wall Street Journal's* story titled "AstraZeneca drug used to fight cancer is tied to 124 deaths": "Side effects from the cancer-fighting drug Iressa have resulted in 124 deaths in Japan, a government official here said, as a ministry panel set stricter guidelines for the drug's use. .... Early studies showed lung-cancer patients who hadn't been helped by other therapies recovered impressively after taking Iressa [the impressive recovery refers to the study above where only 10% of patients had tumor shrinkage], but the large number of severe side effects prompted AstraZeneca to issue a safety warning to Japanese doctors in October".

Despite the report of deaths in Japan, the U.S. Food and Drug Administration (FDA) approved *Iressa*.[6] In fact, this was their first accelerated approval –a new program to more rapidly approve drugs for dangerous diseases with limited treatment options– like lung cancer –based on early studies. Accelerated approval is typically conditional upon confirmation of the results in a randomized trial post-approval. The *Wall Street Journal* was extremely enthusiastic about this decision. Their editorial board wrote: "A rare victory at the FDA. When an FDA advisory panel convened Tuesday to consider AstraZeneca's application for the cancer drug Iressa, it was expected to send the company back for more data. But spurred on by powerful testimony from patients who would almost surely be dead without the drug, and over the apparent objections of hypercautious FDA staffers, the panel voted 11-3 to recommend Iressa for accelerated approval".

Hope was even higher when another study was published in *The New England Journal of Medicine*.[7] This study received hyperbolic coverage –particularly on the U.S. national news. NBC national news ran a segment "Scientists announce major breakthrough in treatment of lung cancer with *Iressa*". The segment featured the story of a young woman with children whose tumor melted away and quotes experts who say the drug will save thousands of lives. Surprisingly, the *The New England Journal of Medicine* article was reporting on 16 of the patients from the original *Iressa* study (where all of the patients received *Iressa*). The "new" study found that eight of the nine patients who responded to Iressa (experienced tumor shrinkage) had a specific genetic mutation while none of the seven patients who had not responded to *Iressa* lacked the mutation. But since all patients with the mutation received Iressa, there is (as described above) no way to know if the mutation predicted response to the drug. Unfortunately, subsequent work failed to confirm that the genetic mutation predicted response to this class of drugs.

Even worse, the phase III study (required by the FDA as part of the accelerated approval program) did not find any survival benefit from *Iressa*. In this randomized trial of 1,700 lung cancer patients, the *Iressa* group had a median survival of 5.6 months vs. 5.1 months in the placebo group.[8] The FDA then pulled the drug from the market, only allowing it for compassionate use.[9] Figure 1 summarizes the *Iressa* and the news cycle.

"Major cancer breakthrough? New drug potential 'holy grail' for treatment of cancer"
CBS Healthwatch, The Early Show

Excessive hope about cancer drugs did not end with Iressa story. A major breakthrough in 2009 started with a *The New England Journal of Medicine* article about "Parp inhibitors" (drugs which inhibit poly (ADP ribose) polymerase) in cancer patients who had BRCA mutations.[10] It is hard to exaggerate the exaggeration of the three major national television news networks.[11] In addition to the *CBS* story, *NBC Nightly* news reported "Now we turn to what some are calling the most important cancer treatment breakthrough in a decade", and *ABC* news "New hope: cancer treatment".

To distinguish hope from hype, we need to understand the science behind the headlines. This study measured what happened to 19 patients with BRCA1 or BRCA2 mutations with ovarian, breast or prostate cancer. After about 5 months of follow-up, 63% had either stable disease (stable tumor markers) or response (defined as 30% or more tumor shrinkage on x-ray). This study was a phase I study –an uncontrolled study

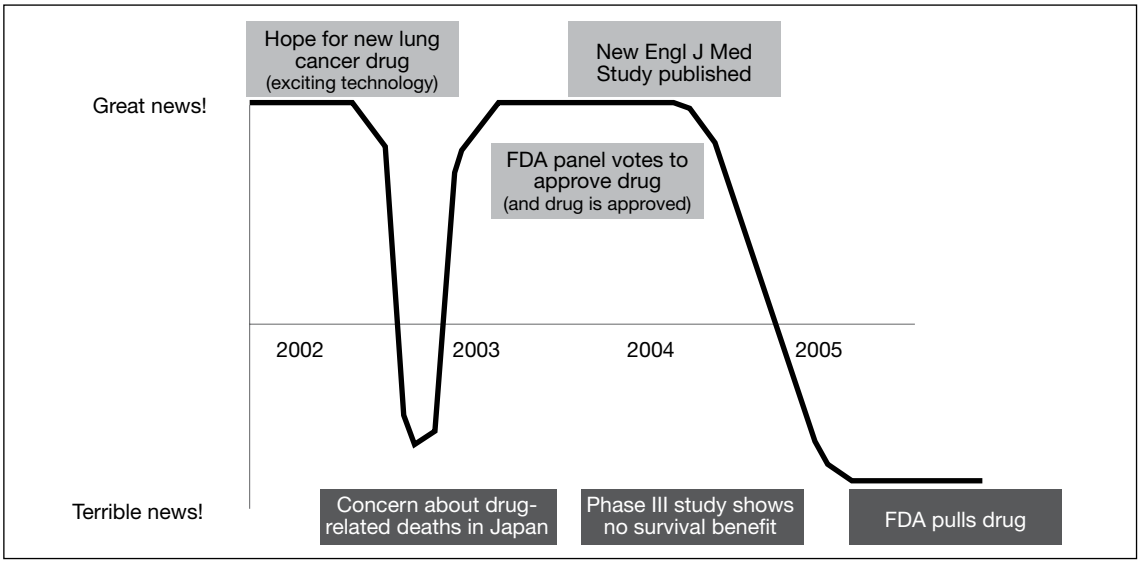


Figure 1.

using a surrogate outcome– just like the *Iressa* study. Publication in *The New England Journal of Medicine* seemed to trump the weakness of the science (which many journalists recognized). In fact, some asserted that since *The New England Journal of Medicine* typically does not publish this kind of uncontrolled study it must mean that this study was very important. Sadly, once again, the randomized trial did not show a difference in cancer death –and the drug company has abandoned pursuing approval.

The Parp inhibitor story reinforces the lessons of *Iressa* –be extremely cautious about uncontrolled studies and surrogate outcomes. But it also holds another important lesson: recognize pseudo-evidence. Publication in a medical journal –even *The New England Journal of Medicine*– does not guarantee the findings are true (or even important). We can all agree that giving false hope to sick patients is a real disservice.

"Your friends may be as powerful as anticancer drugs in the fight against breast cancer" *Women's Day* (magazine), October 2010

"Do you get together with friends often? Here's an important reason to accept your pal's book club invitation: An active social life is not only good for your general health, but keeping up with your girlfriends may also reduce your risk of developing breast cancer. In a recent study, researchers at the University of Chicago report that lonely women may be at greater risk of breast cancer. The theory? Stress and anxiety caused by social isolation may have the power to increase the growth of tumors in the breast."

Is it really possible that all you need to do is spend some time with your friends to reduce your risk of breast cancer? This magazine story is based on a study published in the prestigious journal, *Proceedings of the National Academy of Sciences*.[12]

The story, however, wildly extrapolates findings from a study of 40 rats. Half the rats were randomized to live alone from 1 month of birth until death (the other lived in groups of five-female rats). Rats were just as likely to develop a breast tumor regardless of isolation; however isolated rats developed more and larger tumors.

Does this really mean that your friends are as powerful as "anticancer drugs"? Inbred rats, genetically altered so they are predisposed to develop breast cancer are not typical rats. And, of course, even typical rats are not like typical humans. Nor is total, lifelong isolation the same as refusing your pal's book club invitation or feeling lonely.

If we had to write about this study, here is what we would say: "News only a mutant rat could use (maybe). This study of mutant rats forced to live in total lifelong isolation has no direct meaning on cancer risk for humans or even ordinary rats. Don't get stressed out by this study of stress. And don't feel like you have to change your social behaviors. The level of social isolation in this rat study was far more extreme than any human being could ever experience."

While test-tube and animal studies can be fundamentally important, the problem is claiming imminent relevance to human health. In a systematic review of "high profile" animal studies, it took an average of 14 years to translate the animal research into human testing.[13] And only one-third of animal studies translated into successful interventions in randomized trials of humans. Moving from animals to humans is a slow and uncertain process.

When reporting on such research, extrapolate with caution. Do not tell people what to worry about –or do– based on very preliminary animal or lab science. We recommend cautioning readers that "It takes many years to learn if the findings of animal [lab] studies apply to people. Many promising animal [lab] studies fail to pan out in people".

Sometimes the news is not fit to print. Consider asking your editor whether you can skip it. If you have to report it, always include STRONG cautions. These will help readers avoid over interpreting the findings –and may even sway your editor against covering the study after all.

References

1. Schwartz L, Woloshin S, Baczek L. Media coverage of scientific meetings: too much, too soon? JAMA. 2002;287:2859-63.
2. Comparison of raloxifene hydrochloride and placebo in the treatment of postmenopausal women with osteoporosis (MORE) trial protocol NCT00670319. Available at: http://clinicaltrials.gov/ct2/show/NCT-00670319www.clinicaltrials.gov
3. Cummings S, Eckert S, Krueger K. The effect of raloxifene on risk of breast cancer in postmenopausal women: results from the MORE randomized trial. JAMA. 1999;281:2180-97.
4. Wolosshin S, Schwartz L. What's the rush? The dissemination and adoption of preliminary research results. J Natl Cancer Inst. 2006;98:372-3.
5. Pollack A. Drug's approval hints at flexibility in FDA process. New York Times. May 6, 2003;C; Column 5: 1.
6. United States Food and Drug Administration. Questions and answers on Iressa (gefitinib). 2005. (Accessed January 28, 2006.) Available at: http://www.fda.gov/cder/drug/infopage/iressa/iressaQ&A2005.htm
7. Lynch T, Bell D, Sordella R. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;350:2129-39.
8. Pollack A. Lung-cancer drug shows unfavorable trial results. New York Times. December 20, 2004;C; Column 1: 2.
9. Pollack A. FDA restricts access to cancer drug, citing ineffectiveness. New York Times. June 18, 2005; C; Column 1: 2.
10. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med. 2009;361:123-34.
11. Woloshin S, Schwartz L, Kramer B. Promoting healthy skepticism in the news: helping journalists get it right. J Natl Cancer Inst. 2009;101:1596-9.
12. Hermes G, Delgado B, Tretiakova M. Social isolation dysregulates endocrine and behavioral stress while increasing malignant burden of spontaneous mammary tumors. PNAS. 2009;106:22393-8.
13. Hackam DG, Redelmeier DA. Translation of evidence from animals to humans. JAMA. 2006;296:1731-2.

El azar permite cuantificar la precisión del muestreo

Cobo / Ventura

33 mensajes clave sobre bioestadística para periodistas y comunicadores

Gonzalo Casino

1. *Tipo de estudio*
   «Un estudio» es demasiado vago. Hay que informar de los detalles del estudio y de la confianza que merece.

2. *Riesgo absoluto*
   El riesgo relativo es más aparatoso, pero no hay que olvidarse del absoluto, que pone en perspectiva una amenaza –o un beneficio– para la salud.

3. *Riesgos de la prevención*
   La prevención tiene también sus perjuicios y sus excesos. En las estadísticas del cribado, la supervivencia no es el reverso de la mortalidad.

4. *Divulgar*
   La estadística puede ser compleja, pero tenemos que hacerla sencilla. Hay que traducir la terminología y los números.

5. *Fuentes*
   Atención a las exageraciones de los intermediarios (comunicados de prensa). Conviene acudir a las fuentes originales (artículo científico) y contrastar con fuentes independientes y competentes en bioestadística.

6. *Contexto*
   Un estudio es una frase anecdótica en medio de una conversación. Lo que interesa es la conversación completa.

Erik Cobo

7. *Variabilidad*
   La estadística aborda la variabilidad: su obsesión es cuantificar la incertidumbre (otro tema es si los usuarios entran...)

8. *Predecir y modificar*
   No hay que confundir la predicción del futuro

con su modificación, que requiere relación causal.

9. *Reducción de la incertidumbre*
   Rara vez una predicción anula toda duda; por eso debe acompañarse de medidas de reducción de la incertidumbre.

10. *Estudio experimentales*
    La clave de un estudio experimental es la asignación: sólo variables que dependen del investigador pueden cambiarse y así estimar sus efectos.

11. *Causas y efectos*
    Se empieza por buscar causas y se termina por estimar efectos.

12. *Ciencia y técnica*
    La pregunta de la ciencia es «qué sé» y la de la técnica es «qué hago». La primera descansa en la evidencia, pero la segunda debe contemplar también las consecuencias.

13. *La p y el intervalo de confianza*
    Es más importante el intervalo de confianza que el valor de p. Si un resultado no lleva un intervalo de confianza relevante, mejor ignorarlo.

14. *Objetivos sanitarios*
    Cada objetivo sanitario (tratamiento, pronóstico, diagnóstico) tiene un tipo de estudio adecuado (ensayo clínico, cohorte, transversal).

Pablo Alonso

15. *Calidad*
    La calidad es la confianza/certidumbre que tenemos en que los resultados obtenidos provenientes de la investigación sean ciertos.

16. *Revisiones sistemáticas*
    Para contextualizar y conocer con mayor seguridad el efecto de las intervenciones son clave las revisiones sistemáticas (con o sin metaanálisis).

17. *Diseño*
    Para conocer la calidad/confianza, el diseño de los estudios es importante, pero no suficiente.

18. *Menos confianza*
    Según el sistema GRADE, la confianza en los ensayos clínicos (inicialmente considerada como alta) puede disminuir cuando hay:
    - Riesgo de sesgo (aleatorización/cegamiento/pérdidas).
    - Inconsistencia, imprecisión, evidencia indirecta o sesgo de publicación.

19. *Más confianza*
    Según el sistema GRADE, la confianza en los estudios observacionales (de entrada baja) puede aumentar si hay:
    - Asociación importante o gradiente dosis-respuesta.
    - Cambio radical de pronóstico o inmediatez del efecto.

José Luis Peñalvo

20. *Carga*
    La epidemiología utiliza medidas de “carga” para caracterizar la enfermedad y proponer hipótesis causales. Estas hipótesis se estudian, en la mayoría de los casos, mediante comparación de grupos con las denominadas medidas de riesgo y de asociación: riesgo relativo (RR) y *odds ratio* (OR), que explican cuánto más riesgo tiene el grupo expuesto en comparación con el no expuesto.

21. *Cohorte*
    Estudio observacional longitudinal y prospectivo. Clasificación de la población según exposición y espera hasta suceso de interés. Medida de asociación: riesgo relativo (RR).

22. *Casos y controles*
    Estudio observacional longitudinal y retrospectivo. Clasificación de la población según sucesos y recuperación de información sobre exposición. Medida de asociación: *odds ratio* (OR).

23. *Causalidad*
    Relación etiológica entre una exposición y la aparición de un efecto. Modelo de Bradford-Hill (1965), que propone los siguientes criterios de causalidad: fuerza de la asociación, consistencia, especificidad, temporalidad,

gradiente biológico, plausibilidad biológica, coherencia, evidencia experimental y analogía.

24. *Confusión*
    Tercera variable que "confunde" total o parcialmente la relación entre la exposición y el efecto, y no forma parte del mecanismo causal. En los estudios observacionales se utiliza la estratificación o el ajuste de los modelos (análisis multivariado) para mejorar la asociación.

Steven Woloshin

25. *Sources exaggerate*
    Medical journals, academic medical centers and researchers all contribute to the problems with health news.

26. *Too much too soon*
    News stories about scientific meeting presentations often lack basic facts, numbers and cautions.

27. *Recognize disease mongering campaigns*
    Be skeptical when new –or expanded– diseases are being promoted:
    - Question prevalence estimates.
    - Present the full spectrum of disease.
    - Question idea that more diagnosis is always better.
    - Quantify the benefits and harms of the new treatment.

28. *If you take away nothing else...*
    - Use numbers!
    - Highlight cautions!

Lisa Schwartz

29. *Scientific meeting research*
    These preliminary findings may change because the study has not been independently vetted through peer review and all the data are not yet in.

30. *No control group*
    Because everyone took drug, it is extremely hard to know if drug had anything to do with the outcome.

31. *Surrogate outcomes*
    This study measured [surrogate] –a lab test [or x-ray] finding– that people do not directly experience. Be cautious about acting on these findings since changes in these measures do not reliably translate into people feeling better or living longer.

32. *Recognize pseudo-evidence*
    Publication in a medical journal –even *The New England Journal of Medicine*– does not guarantee the findings are true (or even important).

33. *Animal or lab study*
    It takes many years to learn if the findings of animal [lab] studies apply to people. Many promising animal [lab] studies fail to pan out in people. Extrapolate with caution! don't tell people what to worry about –or do– based on very preliminary animal or lab science.

La variabilidad, esa amiga de la vida

Cobo / Ventura

Bibliografía

Bibliografía recomendada

- Alonso-Coello P, Rigau D, Solà I, Martinez García L. La formulación de recomendaciones en salud: el sistema GRADE. Med Clin (Barc). 2013;140:366-73.
- Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. El significado de las estadísticas. Mente y cerebro. 2011;50: 62-9.
- Gisbert JP, Bonfill X. Cómo realizar, evaluar y utilizar revisiones sistemáticas y metaanálisis. Gastroenterol Hepatol. 2004;27:129-49.
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924-6.
- Interpretando la literatura médica: ¿qué necesito saber? Parte I. Disponible en: http://www.osakidetza.euskadi.net/r85-pkcevi04/eu/contenidos/informacion/cevime_infac/eu_miez/adjuntos/infac_v14_n7.pdf
- Interpretando la literatura médica: ¿qué necesito saber? Parte II. Disponible en: http://www.osakidetza.euskadi.net/r85-ckpubl01/eu/contenidos/informacion/cevime_infac/eu_miez/adjuntos/infac_v14_n8.pdf
- Woloshin S, Schwartz LM, Welch HG, editores. Know your chances. Understanding health statistics. Berkeley (CA): University of California Press; 2008. Disponible en: www.ncbi.nlm.nih.gov/pubmedhealth/n/ucalpkyc/pdf/

Otras fuentes

- Colaboración Cochrane (www.cochrane.org). Organización sin ánimo de lucro que agrupa a miles de investigadores de 90 países para la realización de revisiones sistemáticas de los estudios disponibles sobre las intervenciones en salud.
- PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). Buscador de la base de datos Medline de la National Library of Medicine de Estados Unidos, que ofrece referencias y resúmenes de las principales publicaciones de biomedicina de todo el mundo desde 1966.
- Red CASPE (http://redcaspe.org/drupal/?q=node/29). Herramientas para el análisis crítico de la literatura científica.
- TripDatabase (http://www.tripdatabase.com). Metabuscador de ensayos clínicos, revisiones sistemáticas y otras investigaciones, que ofrece los resultados de la búsqueda categorizados.

La ciencia acompaña sus afirmaciones
de medidas de incertidumbre

Ventura

Apéndice

Numbers glossary

All examples are based on the following scenario:

In a randomized trial, 200 adults were given either DRUG or placebo for 5 years. Here's what happened:

	EXPOSED DRUG (100 adults)	CONTROL Placebo (100 adults)
Died during study	10 people	30 people

Measure	Definition	Example
Absolute risk Analogy: Price Absolute risk (*control*) is the regular price. Absolute risk (*exposed*) is the *sales* price.	Number who had outcome / Number who could have had outcome	**Absolute risk** (DRUG group) = 10/100=0.10=**10%** **Absolute risk** (Placebo group) = 30/100=0.30=**30%** Over 5 years, **10%** of the DRUG group died compared to **30%** of the placebo group. DRUG lowered the chance of dying compared to placebo: **10%** vs. **30%** died over 5 years.
Absolute risk reduction (ARR) "percentage points lower" Analogy: Savings from a sale. Subtract the sales price from the regular price.	Absolute risk (control) – Absolute risk (exposed)	**Absolute risk reduction** = 30% - 10% = **20% = 20 in 100** **For every 100 people** who take DRUG instead of placebo for 5 years, **20 fewer** people would die. DRUG lowered the chance of dying over 5 years by **20 percentage points** compared to placebo: 10% vs 30%.
Number needed to treat (NNT)	1 / Absolute risk reduction	**Number needed to treat** = 1 / 20% = 1/0.20 = **5** **5 adults** would have to take DRUG for five years to prevent **1 death**.
Relative risk (RR)	Absolute risk (exposed) / Absolute risk (control)	**Relative Risk** = 10% / 30% = 0.1/0.3 = **0.33** The DRUG group had **0.33** times the chance of dying compared to placebo: 10% vs 30% died over 5 years. The DRUG group had **one third** the deaths of the placebo group: 10% vs 30% died over 5 years.
Relative risk reduction (RRR) "% lower" Analogy: "% off" for the sale ("67% off regular price")	1 – Relative Risk	**Relative risk reduction** = 1 - 0.33 = 0.67 or **67%** DRUG reduced the chance of dying by **67%** compared to placebo: 10% vs 30% died over 5 years. DRUG lowered deaths by **two thirds** compared to placebo: 10% vs 30% died over 5 years.

Bottom Line Always report absolute risks for each group (no matter what other numbers are used).

For all risks, you need to be clear about 3 things: exactly what the outcome is (e.g. having a heart attack), over what time period the outcome occurred (e.g. 5 years) and in whom (e.g. adults with diabetes).

Steven Woloshin and Lisa Schwartz.
Center for Medicine and the Media, Dartmouth Institute for Health Policy and Clinical Practice.

Statistics glossary

The p value and confidence interval are based on the following scenario

In a randomized trial, 200 adults were given either DRUG or placebo for 5 years. Here's what happened:

	EXPOSED DRUG (100 adults)	CONTROL Placebo (100 adults)
Died during study	10 people	30 people

Measure	Definition	Example
p value	Probability that an observed effect size is due to chance *alone* if $p > 0.05$, we say "likely due to chance", "not statistically significant" if $p < 0.05$, we say "unlikely due to chance", "statistically significant" Remember, even with a very low p value ("highly statistically significant"), results can still be very wrong: the study may be biased or confounded.	Relative risk reduction = 67%, **p=0.0004** The observed differences in the 5 year risk of death between the DRUG and placebo group is not consistent with chance alone (i.e. p=0.0004–there is only a **4 in 10,000** chance of seeing differences this big or bigger if DRUG and placebo were the same). These results are very unlikely to be due to chance.
Confidence interval (95% CI)	Because the observed value is only an estimate of the truth, we know it has a "margin of error". The range of plausible values around the observed value that will contain the truth 95% of the time.	Relative risk reduction **(95% CI) =** 67%(**36%– 83%**) While our best estimate is that DRUG lowers the 5 year risk of death by 67%, the results of this study say it is possible that DRUG may lower the risk **by as little as 36% or as much as 83%**
Early Detection Statistics		
Survival	$\frac{\text{Number alive at a specfied time after Cancer X diagnosis (typically 5 or 10 years)}}{\text{Number diagnosed with Cancer X}}$ Comparing survival of patients diagnosed by different methods tells you nothing about the benefit of early detection. Consequently, comparing survival across time (e.g. 1970 vs. 2008) or place (e.g. UK vs. US) – when patterns of testing are different – is misleading. They cannot tell you whether anyone is living longer.	10-year lung cancer survival was: **29%** for patients diagnosed by screening chest x-rays **14%** for patients diagnosed by symptoms Lung cancer patients diagnosed by screening chest x-rays have a 10-year survival of 29% compared to **14%** of lung cancer patients diagnosed by symptoms, like cough or weight loss. Warning: This statement is misleading. It tells you nothing about about the benefit of screening.
Mortality	$\frac{\text{Number of Cancer X deaths over a specified time}}{\text{Total No. of people in study or population (i.e. with \& without Cancer X diagnosis)}}$ Reduced mortality in a randomized trial is the only reliable evidence for the benefit of screening.	In a randomized trial of chest x-ray screening, **10 year lung cancer mortality** was: **4%** for the chest x-ray screening group **4%** for the control group (not screened) The **10-year lung cancer mortality** among the chest x-ray screening group was **4%** versus **4%** in the control group.

THE DARTMOUTH INSTITUTE FOR HEALTH POLICY & CLINICAL PRACTICE

Woloshin & Schwartz
VA Outcomes Group (http://www.vaoutcomes.org)
Center for Medicine and the Media

Steven Woloshin and Lisa Schwartz.
Center for Medicine and the Media, Dartmouth Institute for Health Policy and Clinical Practice.

Questions to guide your reporting

What is the finding?

What is the distinct exposure – or treatment – in each group?

If it is a lifestyle exposure (diet or exercise), how does it translate into what you have to eat or do?

What is the outcome under consideration?

If the outcome is a surrogate (cholesterol test), is it strongly linked to patient outcomes (heart attack)?

If the outcome is a composite (combining multiple components such as heart attack, stroke, or death), can you learn about the role of each component?

If the outcome is a score, (Hamilton Rating Scale for Depression) can you learn what constitutes a clinically important difference (that patients can notice) and what proportion in each group experienced it?

How big is the finding?

What is the chance of the outcome (over what time period) in each group?

Just knowing the relative risk ("0.75 times the risk") or the relative risk reduction ("25% fewer") without knowing the absolute risk is insufficient. Remember that a relative risk of 0.75 can represent an infinite number of combinations (0.003% vs. 0.004%, 3% vs. 4%, 30% vs. 40%)

Guidelines for presenting absolute risks

- Present absolute risks for each exposure group along with the time frame.
- Consider expressing absolute risks as percents (10%). This format is understandable even for decimal percents (0.5%).
- If expressing absolute risks as frequencies, DON'T use the "1 in X" format which makes comparisons hard (1 in 35 vs 1 in 56).
  Instead use "X in ___" like 2 in 1000.
  For the "in ___" part use multiples of 10, choosing the smallest one which makes "X" a whole number.
  Use the same "in ___" for the whole story.
- Provide context for the absolute risk.
  How dangerous is the disease? Compare absolute risk of getting to dying from disease.
  How does this risk compare to others? Compare absolute risk of dying from cancer to dying from heart disease.

What are the downsides of intervention: life threatening harms, bothersome side effects, inconvenience, costs?

When reporting on a beneficial treatment, make sure you look for associated harms. And report the absolute risks for these harms in the same format, for the same time frame, and the same dose.

Special case: Odds ratios overstate effects when outcomes are common (when the absolute risk is >20%).

Always ask: What are the absolute risks in each exposure group?

What does the finding mean?

How does the finding fit with what is already known about the topic?

Look for a systematic review.

Is the finding clinically meaningful or just "statistically significant" (i.e., p<0.05)?

Is the outcome something people directly experience or really care about? Is the effect size big or small?

Avoid the word significant. Consider using "important" for clinically meaningful and "unlikely to be due to chance" for statistical signficance.

Could the finding be wrong?

If an observational study, consider how likely it is that confounding -- differences between the people in the exposure groups — might explain the finding?

How different are the exposure groups in terms of age, sex, income, illness level, behaviors like smoking?

Did the investigators attempt to deal with confounding? How much did adjustment change the finding?

If a negative study (effect size not statistically significant), ask whether the confidence interval includes a clinically meaningful effect?

Special case: 5-year survival and screening

Improved 5-year survival for screened vs unscreened patients tells you nothing about the benefit of screening.

Bottom Line

If you can't get answers, consider skipping the story. Use numbers (and put them in tables) and highlight cautions.

Steven Woloshin and Lisa Schwartz.
Center for Medicine and the Media, Dartmouth Institute for Health Policy and Clinical Practice.

How to highlight study cautions

Setting	Suggested Language
Preliminary research (unpublished scientific meeting presentations)	These preliminary findings may change because the study has not been independently vetted through peer review [and/or] all the data are not in yet.
Inherently weak designs Animal or lab study	It takes many years to learn if the findings of animal [or lab] studies apply to people. Many promising animal [lab] studies fail to pan out in people.
Cross sectional study	Because all information was collected at the same time, you can't know if [exposure] caused [outcome], or visa versa.
Ecological studies (International comparison of dietary fat consumption vs. colon cancer mortality rate)	The study provides weak evidence connecting [exposure] and [outcome]. It shows that populations with more [exposure] have more/less [outcome]. But the study cannot tell if the people [with exposure] are the ones who actually had the [outcome].
Models (decision analysis)	The findings are based on assumptions including hypothetical relationships which may not exist.
No control group	Because everyone [took the drug / had the exposure], it is extremely hard to know if the [drug/exposure] had anything to do with the outcome.
Small study (less than 30 people)	These findings are based on a small study; larger studies are needed to really understand how much the intervention works.
Surrogate outcomes (lab test or x-ray finding)	This study measured [surrogate outcome] - a lab test/ x-ray finding - that patients do not directly experience. Be cautious about acting on these findings since changes in these kinds of measures do not reliably translate into people feeling better or living longer.
Classic designs Randomized trial	
Extrapolation	The findings may not apply to people who differ from those in the study (people with less severe disease or at lower risk for bad outcomes).
New interventions	The study only lasted a short time - [X days, weeks or months]. The balance of benefits and harms may change over a longer time period. Longer-term studies are needed.
New drugs	[Drug] is new: it was approved in [year]. As with all new drugs, we don't know how its safety record will hold up over time. In general, if there are unforeseen, serious side effects, they emerge after the drug is on the market when a large enough number of people have used the drug.
Observational studies (with a control group)	
Trial not possible (harmful exposure)	Because the study was not a true experiment, the findings may be explained by differences in the people who happened to be [exposed] rather than [drug/exposure].
Trial possible (beneficial exposure)	Because the study was not a true experiment, we cannot know whether changing [exposure] will change [outcome]. The findings may be explained by differences in the people who happened to be [exposed] rather than [drug/exposure]. A randomized trial is needed before widespread adoption of [intervention].
All studies	The benefit of [any action/intervention] should be weighed against the [side effects, inconveniences, costs, etc.].

THE DARTMOUTH INSTITUTE FOR HEALTH POLICY & CLINICAL PRACTICE

Woloshin & Schwartz
VA Outcomes Group
(http://www.vaoutcomes.org)
Center for Medicine and the Media

Bottom Line

Use cautions – all studies have them.
Consider not reporting preliminary or inherently weak research.

Steven Woloshin and Lisa Schwartz.
Center for Medicine and the Media, Dartmouth Institute for Health Policy and Clinical Practice.

CUADERNOS DE LA FUNDACIÓN DR. ANTONIO ESTEVE

1. Guardiola E, Baños JE. Eponímia mèdica catalana. Quaderns de la Fundació Dr. Antoni Esteve, Nº 1. Barcelona: Prous Science; 2003.
2. Debates sobre periodismo científico. A propósito de la secuenciación del genoma humano: interacción de ciencia y periodismo. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 2. Barcelona: Prous Science; 2004.
3. Palomo L, Pastor R, coord. Terapias no farmacológicas en atención primaria. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 3. Barcelona: Prous Science; 2004.
4. Debates sobre periodismo científico. En torno a la cobertura científica del SARS. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 4. Barcelona: Prous Science; 2006.
5. Cantillon P, Hutchinson L, Wood D, coord. Aprendizaje y docencia en medicina. Traducción al español de una serie publicada en el British Medical Journal. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 5. Barcelona: Prous Science; 2006.
6. Bertomeu Sánchez JR, Nieto-Galán A, coord. Entre la ciencia y el crimen: Mateu Orfila y la toxicología en el siglo XIX. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 6. Barcelona: Prous Science; 2006.
7. De Semir V, Morales P, coord. Jornada sobre periodismo biomédico. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 7. Barcelona: Prous Science; 2006.
8. Blanch Ll, Gómez de la Cámara A, coord. Jornada sobre investigación en el ámbito clínico. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 8. Barcelona: Prous Science; 2006.
9. Mabrouki K, Bosch F, coord. Redacción científica en biomedicina: Lo que hay que saber. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 9. Barcelona: Prous Science; 2007.
10. Algorta J, Loza M, Luque A, coord. Reflexiones sobre la formación en investigación y desarrollo de medicamentos. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 10. Barcelona: Prous Science; 2007.
11. La ciencia en los medios de comunicación. 25 años de contribuciones de Vladimir de Semir. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 11. Barcelona: Fundación Dr. Antonio Esteve; 2007.
12. Debates sobre periodismo científico. Expectativas y desencantos acerca de la clonación terapéutica. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 12. Barcelona: Fundación Dr. Antonio Esteve; 2007.
13. Gonzàlez-Duarte R, coord. Doce mujeres en la biomedicina del siglo XX. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 13. Barcelona: Fundación Dr. Antonio Esteve; 2007.
14. Mayor Serrano MB. Cómo elaborar folletos de salud destinados a los pacientes. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 14. Barcelona: Fundación Dr. Antonio Esteve; 2008.
15. Rosich L, Bosch F, coord. Redacció científica en biomedicina: El que cal saber-ne. Quaderns de la Fundació Dr. Antoni Esteve, Nº 15. Barcelona: Fundació Dr. Antoni Esteve; 2008.
16. El enfermo como sujeto activo en la terapéutica. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 16. Barcelona: Fundación Dr. Antonio Esteve; 2008.
17. Rico-Villademoros F, Alfaro V, coord. La redacción médica como profesión. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 17. Barcelona: Fundación Dr. Antonio Esteve; 2009.
18. Del Villar Ruiz de la Torre JA, Melo Herráiz E. Guía de plantas medicinales del Magreb. Establecimiento de una conexión intercultural. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 18. Barcelona: Fundación Dr. Antonio Esteve; 2009.
19. Gonzàlez-Duarte R, coord. Dotze dones en la biomedicina del segle XX. Quaderns de la Fundació Dr. Antoni Esteve, Nº 19. Barcelona: Fundació Dr. Antoni Esteve; 2009.
20. Serés E, Rosich L, Bosch F, coord. Presentaciones orales en biomedicina. Aspectos a tener en cuenta para mejorar la comunicación. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 20. Barcelona: Fundación Dr. Antonio Esteve; 2010.
21. Francescutti LP. La información científica en los telediarios españoles. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 21. Barcelona: Fundación Dr. Antonio Esteve; 2010.

22. Guardiola E, Baños JE. Eponímia mèdica catalana (II). Quaderns de la Fundació Dr. Antoni Esteve, Nº 22. Barcelona: Fundació Dr. Antoni Esteve; 2011.
23. Mugüerza P. Manual de traducción inglés-español de protocolos de ensayos clínicos. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 23. Barcelona: Fundación Dr. Antonio Esteve; 2012.
24. Marušić A, Marcovitch H. Competing interests in biomedical publications. Main guidelines and selected articles. Esteve Foundation Notebooks, Nº 24. Barcelona: Esteve Foundation; 2012.
25. De Semir V, Revuelta G, coord. El periodismo biomédico en la era 2.0. Cuadernos de la Fundación Dr. Antonio Esteve, Nº 25. Barcelona: Fundación Dr. Antonio Esteve; 2012.

Para solicitar cuadernos ya publicados diríjanse por escrito a la Fundación Dr. Antonio Esteve, c/ Llobet i Vall-Llosera nº2, 08032 Barcelona, o a través de la página web: www.esteve.org.