2

Causalidad y asociación estadística

M. Porta Serra y E. Fernández Muñoz

□ INTRODUCCIÓN

La aproximación al estudio patogénico de una enfermedad o un síndrome parte de un modelo teórico —implícito o explícito— de causalidad. La clasificación y el diagnóstico diferencial de las enfermedades son procesos que están asimismo condicionados por la elección de un modelo causal. A continuación se presentan en forma resumida algunos elementos esenciales del marco teórico en el cual se ha planteado el concepto de causalidad en medicina, y sus implicaciones para la medicina clínica y la epidemiología.

■ CAUSALIDAD

Modelo determinista puro

La utilización de una metodología «epidemiológica» para establecer la causa y/o el modo de transmisión de las epidemias se remonta a la era prebacteriana. Por ejemplo, Snow estudió las epidemias de cólera de Londres de los años 1849 y 1854 y las atribuyó a la acción de un «veneno colérico» transmitido a través del agua contaminada por la red abastecedora de la ciudad; aunque desconocía el agente etiológico de la enfermedad (vibrión colérico), las medidas que tomó para controlar las epidemias fueron eficaces.

El desarrollo de la bacteriología y de sus métodos permitió identificar diferentes agentes causales de diversas enfermedades, aunque las guías para establecer tal causalidad eran casi inexistentes. Los estudios bacteriológicos de Henle, Klebs y los suyos propios, llevaron a Koch a establecer en 1882 unos postulados de causalidad que se conocen como «postulados de Henle-Koch». Él y muchos otros los aplicaron para identificar los agentes causales de las enfermedades y pueden resumirse de la manera siguiente:

1. El agente debe estar presente en cada caso de la enfermedad y en circunstancias que puedan explicar los cambios patológicos y el curso clínico de ésta.
2. El agente no debe estar presente en ninguna otra enfermedad como acontecimiento fortuito y no patogénico.
3. El agente, después de ser aislado a partir de un organismo que tiene la enfermedad y crecer de manera continuada en cultivo puro, debe ser capaz de inducir de nuevo la enfermedad.

Este modelo, que tantos progresos aportó a la medicina durante muchos años, se convirtió después en un lastre notable para la investigación médica, sobre todo cuando se aplicó a nuevos agentes infecciosos (p. ej., virus) o a otros grupos de trastornos (enfermedades degenerativas,

neoplásicas, inmunológicas, tóxicas). De hecho, el propio Koch se dio cuenta de que algunos agentes etiológicos no seguían fehacientemente sus tres postulados y consideró que el cumplimiento de los dos primeros podía ser suficiente demostración de causalidad.

Casi un siglo más tarde se reconocieron las principales limitaciones de este modelo de causalidad tanto para las enfermedades agudas como para las crónicas:

1. La misma situación patológica o clínica puede ser resultado de diferentes factores de riesgo en distintas circunstancias.
2. Los factores de riesgo pueden variar según las áreas geográficas, los grupos de edad, los comportamientos y los patrones de susceptibilidad (sobre todo, genéticos).
3. En la mayoría de las enfermedades, incluyendo las infecciosas, deben concurrir varios factores de riesgo.
4. Un mismo factor de riesgo o un mismo agente pueden inducir diferentes respuestas clínicas y patológicas en diferentes circunstancias.
5. La respuesta de un organismo a uno o varios factores de riesgo puede variar desde la enfermedad inaparente o subclínica hasta la enfermedad grave y fatal.

Sobre la base de estas consideraciones se ha propuesto una serie de guías para el establecimiento de la causalidad, conocidas como «postulados de Evans» (tabla 2-1).

Tabla 2-1 ***Postulados de Evans***

1. La prevalencia de la enfermedad debería ser significativamente más elevada entre los individuos expuestos a la hipotética causa que entre los controles no expuestos
2. La exposición a la hipotética causa debería ser más frecuente entre quienes tienen la enfermedad que entre los controles que no la tienen, permaneciendo constantes todos los demás factores de riesgo
3. La incidencia de la enfermedad debería ser significativamente más elevada entre los expuestos a la hipotética causa que entre los no expuestos, a través de datos conocidos mediante estudios prospectivos
4. La enfermedad debería seguir a la exposición al hipotético agente causal mostrando una distribución en campana, de acuerdo con el período de incubación
5. El espectro de respuestas del huésped debería seguir a la exposición mostrando un gradiente lógico biológico, desde las manifestaciones leves hasta las graves
6. Después de la exposición a la hipotética causa debería ser muy probable la aparición de una respuesta cuantificable del huésped (anticuerpos, células cancerosas) entre aquellos que no la presentaban antes de la exposición, o debería incrementarse si ya estaba presente antes de la exposición. Este patrón de respuesta debería ocurrir de manera infrecuente entre las personas no expuestas
7. La reproducción experimental de la enfermedad debería ocurrir con mayor frecuencia entre los animales o los seres humanos adecuadamente expuestos a la hipotética causa que entre los no expuestos; dicha exposición puede ser intencional en voluntarios, inducida experimentalmente en el laboratorio o reflejar una exposición natural
8. La eliminación o modificación de la hipotética causa debería disminuir la incidencia de la enfermedad (atenuación de un virus, eliminación del alquitrán de los cigarrillos)
9. La prevención de la exposición o modificación de la respuesta del huésped frente a la hipotética causa debería disminuir o eliminar la enfermedad (inmunización, fármacos para reducir el colesterol, factor de transferencia específico linfocitario en el cáncer)
10. Todas las relaciones y los hallazgos deberían tener sentido tanto biológico como epidemiológico

Modelo determinista modificado

Las limitaciones básicas del modelo determinista puro para derivar criterios operativos en la investigación etiológica y en la intervención sociosanitaria pueden sintetizarse en los siguientes puntos: *a)* etiología multifactorial: un enorme número de enfermedades tienen más de una causa; *b)* multiplicidad de efectos: la mayoría de los factores pueden tener más de un efecto patológico, y *c)* el modelo determinista puro no puede definir claramente el papel de factores causales con efectos continuados y persistentes y para los cuales no es posible definir límites precisos de normalidad.

Estas apreciaciones han implicado un cambio en la manera de entender los procesos causales en medicina, con el consiguiente cambio en la terminología; así, el término «causa» se sustituye a menudo por «factor de riesgo», entendiendo como tal la variable o factor que, según se cree, está relacionada con la probabilidad de que un individuo padezca determinada enfermedad. Una causa sería «un hecho o acontecimiento o estado de la naturaleza que inicia o permite, solo o en conjunción con otras causas, una secuencia de acontecimientos que resultan en un efecto».

Otro elemento imprescindible para la aparición de un efecto es el elemento temporal. Todo factor de riesgo necesita cierto tiempo para poder desencadenar su efecto. Ese tiempo o período de inducción es específico para cada factor de riesgo, y una enfermedad puede tener diferentes períodos de inducción según cada uno de sus diferentes factores de riesgo.

En teoría, cada causa suficiente estaría formada por diferentes causas componentes, que son las que se pueden investigar mediante estudios etiológicos. A pesar de ello, el uso cotidiano de la palabra «causa» no establece distinciones entre ambas. La mayoría de las causas de interés en medicina son componentes de causas suficientes, pero no son suficientes por sí mismas. Por ejemplo, beber agua contaminada con el vibrión colérico no es suficiente para contraer el cólera, o estar infectado por el virus de la hepatitis B no es suficiente para desarrollar un hepatocarcinoma. La mayoría de las enfermedades tienen más de un grupo de causas suficientes. Según esta visión de la causalidad —que puede denominarse «modelo determinista modificado»— una misma enfermedad tiene diferentes causas suficientes; cada grupo de causas suficientes consta de diferentes causas componentes, y una misma causa componente puede formar parte a la vez de diferentes causas suficientes. En este modelo una causa necesaria es el factor de riesgo sin el cual no puede producirse la enfermedad.

El ejemplo más claro de una causa necesaria lo constituyen las enfermedades infecciosas bacterianas (p. ej., cólera, tuberculosis), entidades clínicas que se definen a partir de su causa necesaria: el agente microbiano.

La relación entre una posible causa y su efecto puede esquematizarse mediante la utilización de una tabla de contingencia (o tabla de 2 × 2) que se expone en la tabla 2-2:

1. Si *x* es una causa suficiente de *y*, entonces ninguno de los expuestos a *x* estará sano y ninguno de los sanos habrá estado expuesto (c = 0).
2. Si *x* es una causa necesaria de *y*, entonces ninguno de los no expuestos estará enfermo y todos los enfermos habrán estado expuestos (b = 0).

3. Si *x* es una causa necesaria y suficiente de *y*, entonces todos los enfermos habrán estado expuestos y ninguno de los sanos habrá estado expuesto (b = 0 y c = 0).
4. Si X no es una causa necesaria ni suficiente, entonces b ≠ 0 y c ≠ 0.

Al analizar con atención estas cuatro posibilidades se observa que en la vida real las relaciones 1 y 3 son extraordinariamente infrecuentes, por no decir inexistentes. Lo más común en la práctica clínica es que: *a)* para algunas enfermedades, X es necesaria pero no suficiente (la posibilidad 2 mencionada en el párrafo anterior); así, por ejemplo, el bacilo de Koch es (por definición) una causa necesaria de la tuberculosis, pero por sí solo no es suficiente para producirla, y *b)* en otras muchas enfermedades, X no es necesaria ni suficiente (posibilidad 4): ciertamente, «influye», es decir, si un paciente está expuesto al factor X, aunque no es posible asegurar con la máxima certeza que tendrá la enfermedad Y, se sabe que tiene más posibilidades de padecerla. La mayoría de los factores de riesgo en las enfermedades cardiovasculares o neoplásicas constituyen ejemplos de esta situación: el tabaco en relación con el cáncer de pulmón o la hipertensión arterial en relación con la cardiopatía isquémica. La cuarta situación de causalidad mencionada lleva a la consideración del enfoque probabilístico, enfoque que probablemente sea el menos limitado de todos.

Un planteamiento de los procesos patológicos mediante esquemas más o menos deterministas quizá pueda resultar didáctico, pero a menudo favorece las visiones más simplistas y reduccionistas de la salud. Muchas actuaciones inútiles del sistema sanitario —e incluso buena parte de la yatrogenia infligida a las personas y a las comunidades— tienen su origen en modelos causales deterministas.

Por el contrario, la apreciación del hecho multicausal favorece el abordaje de los problemas de salud desde diversas perspectivas: puesto que se acepta que existen varias causas, se puede actuar con diversas herramientas (p. ej., no sólo con fármacos sino también a través de cambios en el entorno familiar y profesional, en el estilo y las condiciones de vida, en el medio ambiente, etc.); en definitiva, tanto ejerciendo la medicina curativa como la medicina preventiva, en el ámbito de la salud pública y mediante cualquier otra actuación que propicie una manera de vivir más saludable.

Tabla 2-2 ***Relación causa-efecto: cuatro situaciones posibles***

		Enfermos de *y*	
		Sí	No
Expuestos a *x*	Sí	a	b
	No	c	d

x, posible causa; *y*, posible enfermedad; a, número de personas expuestas a la hipotética causa que tienen la enfermedad; b, número de personas expuestas que no sufren la enfermedad; c, número de personas no expuestas a la hipotética causa que tienen la enfermedad; d, número de personas no expuestas que no sufren la enfermedad.

Rara vez b y c son, ambas, iguales a cero; es decir, rara vez existe una causa necesaria y suficiente a la vez.

■ CAUSALIDAD Y ASOCIACIÓN ESTADÍSTICA

Aunque la evolución de los criterios de causalidad en medicina expuesta anteriormente parezca un tanto teórica, la práctica médica implica casi a diario el desafío de efectuar inferencias causales. Ello ocurre en el ámbito asistencial, a partir de las observaciones que se hacen de los pacientes (anamnesis, exploración física, estudios complementarios) y de los conocimientos del profesional médico sobre las enfermedades. Y también ocurre al leer e interpretar la literatura médica y cuando se formulan preguntas o hipótesis para la investigación clínica. En síntesis, se establecen inferencias causales mediante dos tipos de procesos lógicos: el razonamiento deductivo y el razonamiento inductivo. A través de un razonamiento deductivo, una determinada proposición de orden general se traslada a un orden específico. Así sucede, por ejemplo, cuando se decide administrar salbutamol a un paciente que presenta una crisis asmática. La decisión se basa en un conocimiento previo de que el salbutamol es eficaz en el tratamiento del broncospasmo. Por lo tanto, se deduce que en el caso de ese paciente en particular, la administración de ese fármaco debería asociarse con una mejoría de la crisis asmática. Mediante el razonamiento inductivo, una determinada proposición de orden específico se traslada a un nivel más general. Cuando se realiza un ensayo clínico para evaluar la eficacia de la heparina en dosis bajas para la prevención de la embolia pulmonar en enfermos recién operados se trata de establecer inducciones que puedan generalizarse a una población determinada de pacientes.

Ambos tipos de razonamiento lógico permiten establecer una teoría mediante la cual una determinada exposición contribuye a la aparición o modificación de un determinado desenlace o efecto. Así, por ejemplo, se puede concluir que un factor de riesgo contribuye al desarrollo de una enfermedad, o que cierto fármaco contribuye a mejorar un cuadro clínico. Para poder efectuar tales inferencias es preciso que concurran tres circunstancias fundamentales: *a)* la asociación debe preceder al desenlace, es decir, debe existir una secuencia temporal; *b)* la asociación no debe obedecer a alguna fuente de error sistemático, y *c)* la asociación entre exposición y desenlace debe ser estadísticamente significativa.

Secuencia temporal

La determinación de la secuencia está dada por la direccionalidad inherente a toda observación. Por ejemplo, un estudio de casos y controles —en el cual la exposición se mide retrospectivamente, *a posteriori* del desenlace— entraña una mayor dificultad en el momento de determinar la secuencia temporal que un estudio de cohortes, en el que la exposición se determina al inicio del estudio y se da por sentado que el desenlace ocurrirá con posterioridad.

Error aleatorio y error sistemático

Es importante conocer los distintos tipos de errores que pueden introducirse en la estimación de los parámetros de una población. Puesto que rara vez es posible obtener información sobre la totalidad de la población objetivo (la

población acerca de la cual se quiere formular una conclusión), a menudo se debe utilizar la información sobre un reducido número de personas, que constituyen la población de muestreo. Esta última puede ser idéntica a la población objetivo, aunque a menudo ambas difieren. Si es así, las conclusiones derivadas de la observación de la población de muestreo no podrán hacerse extensivas a la totalidad de los pacientes. En tal caso, se considera que se ha introducido una fuente de error en el estudio. Tales errores pueden ser de dos tipos: aleatorios y sistemáticos.

Error aleatorio

También denominado error de muestreo, el error aleatorio es el que se presenta cuando existe una diferencia entre la estimación obtenida a partir de los datos estudiados y el parámetro que se pretende estudiar. Se atribuye esencialmente a la variabilidad inherente al proceso de muestreo. La presentación de este tipo de error comprometerá la precisión de una estimación, es decir, el grado de acercamiento entre la estimación obtenida a partir de la información estudiada y el parámetro que se desea estimar. Cuanto menor sea este error, mayor será este acercamiento y también la precisión de la estimación.

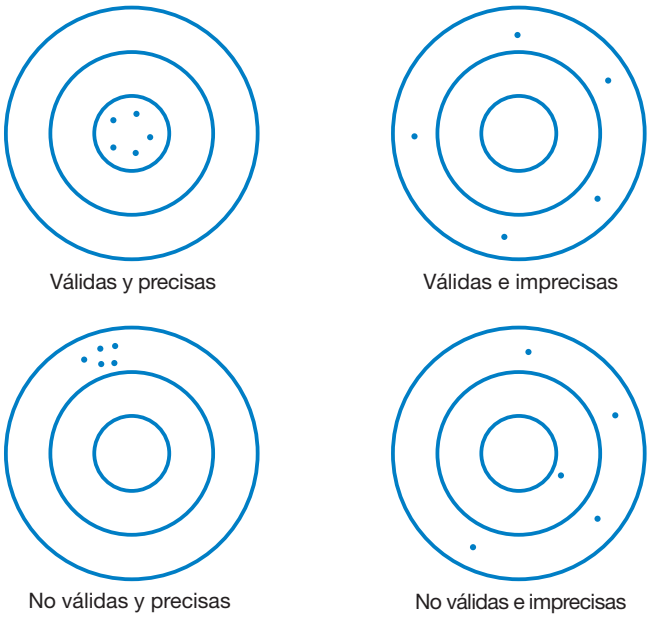

Figura 2-1 ***Analogía de la diana. Validez y precisión de cuatro estimaciones.***

Error sistemático

El error sistemático, también conocido como sesgo, se presenta cuando existe una diferencia entre la estimación y el verdadero efecto que interesa; así, se define «sesgo» como cualquier hecho (estructura del estudio, selección y valoración de los pacientes, metodología de la recolección de datos, análisis de éstos) que hace que los resultados de un estudio se desvíen o aparten de la realidad. Cuanto menores sean los sesgos, mayor será el grado de validez de un estudio.

Los sesgos se pueden clasificar en tres grandes grupos: *a*) los de selección (cuando el resultado obtenido puede explicarse, total o parcialmente, por el modo como se seleccionó a los pacientes del estudio); *b)* los de información (cuando el resultado obtenido puede explicarse, total o parcialmente, porque los grupos de estudio no son similares debido a la forma en que se obtuvieron los datos), y *c)* los factores de confusión (cuando el efecto de la exposición sobre la enfermedad aparece mezclado con los efectos de otro factor de riesgo para la enfermedad que a su vez está asociado a la exposición). La caracterización de estos sesgos en relación con los diferentes tipos de estudios se aborda en los capítulos 4 a 6, y más extensamente en otros textos de metodología.

Para diferenciar los errores aleatorios de los sistemáticos suele usarse la analogía del disparo a una diana (fig. 2-1). La precisión sería semejante a la habilidad para realizar los tiros lo más cerca posible unos de otros, teniendo siempre en cuenta que se intenta acertar en el centro de la diana. Por su parte, la validez se referiría a la habilidad para acertar en el centro de la diana.

Asociación estadística y significación

El tercer componente para poder establecer inferencias es la existencia de una asociación estadísticamente significativa. Considérese el siguiente ejemplo hipotético: se ha diseñado un ensayo clínico aleatorio y a doble ciego para determinar el potencial terapéutico del «respirol», un nuevo dilatador bronquial. En el momento de diseñar el estudio se han tenido en cuenta las consideraciones de validez y precisión.

Ahora se desea evaluar los resultados preliminares, utilizando como medida para el desenlace el número total de crisis asmáticas por paciente, después de 12 meses de administración regular de 30 mg de respirol por día en el grupo tratado; durante este tiempo el grupo de control ha recibido un broncodilatador ya existente en el mercado. El análisis inicial indica que el grupo tratado ha experimentado un 17 % menos de crisis asmáticas por paciente que el grupo de control. Ante esta diferencia cuantitativa entre ambos grupos es preciso formularse la siguiente pregunta: ¿Tiene algún sentido la existencia de tal diferencia entre ambos grupos? El investigador se pregunta si la observación de que los pacientes tratados con «respirol» parecen reaccionar mejor guarda alguna relación con lo que su intuición y experiencia le sugieren. De hecho, puede preguntarse si la diferencia en términos de crisis asmáticas es «significativa». Éste es un término del que a menudo se abusa en las publicaciones médicas. Escribir acerca de la significación de los resultados constituye uno de los «ritos» habituales en las revistas médicas. Sin embargo, puesto que considerar la significación implica establecer una comparación, hace falta un punto de referencia. La pregunta anterior quedará entonces formulada de la siguiente manera: ¿Las diferencias observadas son significativas con respecto a un estándar preestablecido? Ante esta pregunta es posible adoptar, fundamentalmente, dos posturas:

1. En primer lugar, utilizando el sentido común como primer instrumento lógico, el investigador puede preguntarse si las diferencias observadas son significativas con

respecto a lo que cabría esperar en condiciones normales. Lamentablemente, este proceder implica una serie de limitaciones prácticas; entre ellas que, salvo cuando las diferencias fuesen muy grandes, sería difícil convencer a los otros profesionales de que son significativas exclusivamente sobre la base de un criterio subjetivo. Además, el sentido común puede, a veces, jugar malas pasadas. Sin tener en cuenta el número de pacientes estudiados en esta investigación, una diferencia del 17 % en el número de crisis asmáticas entre ambos grupos podría presentarse en proporciones tan dispares como 160/1.000, 114/712 o 4/25.

2. Un segundo elemento de referencia es estadístico. El principal objetivo de la estadística médica es proponer una estrategia gracias a la cual la incertidumbre del proceso de decisión en medicina pueda ser cuantificada (aunque nunca eliminada). Un paso fundamental en muchas investigaciones científicas consiste en formular una hipótesis (la denominada «hipótesis nula») que se tratará de rechazar mediante el uso de una prueba estadística de significación (v. cap. 8), y otra (la «hipótesis alternativa») que se aceptará en caso de que la hipótesis nula sea rechazada. La prueba estadística determinará la fuerza de la evidencia en contra de la hipótesis nula en términos de probabilidad. En el ejemplo, una hipótesis nula que podría ponerse a prueba es: el «respirol» no difiere del broncodilatador habitual en su efectividad clínica en pacientes asmáticos. La hipótesis alternativa sería: el «respirol» difiere del broncodilatador habitual en su efectividad clínica en pacientes asmáticos. La probabilidad de obtener un desenlace por lo menos tan extremo como se habría podido esperar si la hipótesis nula fuera cierta constituye el valor *p* o grado de significación estadística. Cuanto menor sea *p*, más fuerte será la evidencia en contra de la hipótesis nula. Para ayudar al proceso de decisión se establece —arbitrariamente— un nivel de significación estadística (o nivel α), que determinará la cantidad de evidencia en contra de la hipótesis nula que se aceptará como decisiva para rechazarla. Normalmente, este nivel de significación se establece en 0,05 o 0,01. Si *p* es pequeña (inferior al nivel a) significa que la probabilidad de que los resultados obtenidos se deban al azar es pequeña.

El riesgo α (cuantificado como el valor de *p*) indica la probabilidad de cometer un error de tipo I (análogo a un «falso positivo»). Cuanto más pequeño sea el valor de *p*, mayor será la certeza de que la conclusión del ensayo coincide con la «verdad». El riesgo de cometer un error de tipo II se conoce como riesgo β (y es análogo a un «falso negativo»). El objetivo de cualquier estudio es llegar a una conclusión verdaderamente positiva o verdaderamente negativa. La probabilidad de que la conclusión del estudio coincida con la «verdad» será igual a 1 menos beta ($1 - \beta$). Recibe el nombre de potencia o poder estadístico la probabilidad de evitar concluir que no existe un efecto cuando en realidad existe o, lo que es lo mismo, la probabilidad de detectar un efecto si éste existe en verdad (v. cap. 8).

El valor para el riesgo β suele establecerse en el 20 % o, lo que es igual, un poder estadístico del 80 % ($1 - 0{,}20$). Así, si el estudio del ejemplo tiene un poder del 80 %, significa que 8 de cada 10 veces se podrá detectar un efecto, si es que éste existe en realidad, y que 2 de cada 10 veces se asume un riesgo β de de que no pueda llegarse a detectar un efecto real.

¿ESTADÍSTICAMENTE SIGNIFICATIVO O CLÍNICAMENTE IMPORTANTE?

En esta parte final del capítulo se abordará el tema de la significación estadística con dos propósitos específicos. El primero es proporcionar criterios sencillos para orientar al lector de revistas de medicina; tales criterios también serán útiles para el investigador que redacta un artículo. El segundo propósito es muy simple y puede resumirse de la siguiente manera: nunca se debe dar por supuesto que una prueba de significación estadística implica automáticamente la existencia de significación clínica. Por significación clínica se entiende la importancia o relevancia práctica que determinados resultados tienen para el cuidado de los pacientes.

El grado de significación estadística no es una medida de la magnitud del efecto

Un error frecuente consiste en considerar que el valor de *p* es una medida de la magnitud o de la fuerza del efecto o la asociación (p. ej., entre un factor de riesgo y una enfermedad) o del grado de eficacia (p. ej., de un tratamiento). Supóngase que en un estudio sobre la asociación entre las fracturas osteoporóticas y el uso crónico de corticoides se obtienen los siguientes resultados: entre las 200 mujeres estudiadas, habían sufrido fracturas el 30 % de las 100 tratadas con corticoides y el 14 % de las 100 no tratadas (diferencia: 16 %). Antes de realizar el análisis, se había decidido que el nivel de significación estadística sería el 1 %; por lo tanto, se considerarán como estadísticamente significativos aquellos resultados cuyo grado de significación estadística sea inferior al 1 % ($p < 0{,}01$). La prueba de Ji cuadrado da como resultado: $\chi^2 = 7{,}46$ ($p < 0{,}01$) y se podría concluir que en estas mujeres existe una asociación estadísticamente significativa entre el uso crónico de corticoides y el hecho de haber sufrido una fractura osteoporótica. A continuación se analizan los datos de los varones. Entre los 100 varones estudiados, habían sufrido fracturas el 30 % de los 50 tratados con corticoides y el 14 % de los 50 no tratados (diferencia: 16 %). Al aplicar la prueba de Ji cuadrado el resultado es: $\chi^2 = 3{,}73$ ($p > 0{,}05$). Es decir que en estos varones no existe una asociación estadísticamente significativa entre el uso crónico de corticoides y el hecho de haber sufrido una fractura osteoporótica, a pesar de que la diferencia del 16 % es la misma que en las mujeres. En vista de los resultados obtenidos en ambos sexos, podría concluir se que la asociación entre corticoides y fracturas es más fuerte entre las mujeres que entre los varones, ya que la prueba de Ji cuadrado (y, consecuentemente, la significación estadística) es mayor en aquéllas. Sin embargo, la conclusión correcta es que la asociación entre fracturas y corticoides es de la misma magnitud en ambos sexos (16 % de diferencia), y que el número de varones estudiados no fue lo bastante grande como para descartar que la asociación se debiera al azar. ¿Por qué? Porque el grado de significación estadística depende de la magnitud de la diferencia hallada y del número de pacientes analizados, entre otros factores. Por lo tanto, para una misma diferencia (en el ejemplo, 16 % tanto en los varones como en las mujeres), el grado de significación estadística será mayor (*p* menor) cuanto mayor sea el número de pacientes.

Al presentar los resultados de un estudio es importante no dejarse deslumbrar por el grado de significación estadística. Una forma de evitarlo es describir y valorar la magnitud de la diferencia que se ha hallado.

Criterios para la interpretación del grado de significación

La tabla 2-3 presenta diversos puntos que se deben considerar antes y después de llevar a cabo pruebas de significación estadística.

Si *p* es inferior al nivel de significación que se haya establecido previamente (es decir, $p < 0,05$ o bien $p < 0,01$), esto sólo significa que existe una probabilidad pequeña de que los resultados obtenidos se deban al azar. Entonces se dice que la diferencia o la asociación hallada es estadísticamente significativa.

El hecho de que *p* sea «pequeña» (inferior al nivel de significación) puede reflejar dos fenómenos distintos: *a)* que existe una asociación fuerte (o un efecto elevado, o una gran diferencia), lo cual, en principio, es clínicamente importante, y *b)* que el número de pacientes analizados es grande (en un estudio con muchísimos pacientes, algunas diferencias clínica o sanitariamente insignificantes pueden ser significativas desde el punto de vista estadístico).

Tabla 2-3 ***Algunos puntos que se deben considerar antes y después de llevar a cabo pruebas de significación estadística***

Antes

1. Dedique tiempo a revisar críticamente la bibliografía y a pensar lo que a usted, a sus colegas y a los enfermos les interesa realmente saber
2. Formule una o dos hipótesis operativas
3. Escoja el diseño más adecuado y realista para el estudio
4. Defina con claridad uno o dos desenlaces o efectos clínicos de interés
5. Determine concretamente qué variables medirá y qué relaciones entre ellas considerará
6. Elija las pruebas estadísticas que utilizará para ello
7. Decida un nivel de significación estadística, teniendo en cuenta el número de pruebas estadísticas que piensa efectuar cuando disponga de los datos
8. Escoja también un riesgo ß y calcule el poder estadístico (1– ß) que, con un determinado número de pacientes, tendrá para detectar determinadas diferencias en los desenlaces o efectos de interés
9. Recoja únicamente aquellos datos que sean relevantes para poner a prueba su hipótesis, asegurándose de su veracidad

Después

1. No se deje fascinar por la *p*: describa sus resultados con independencia de su significación estadística (mediante tablas y figuras)
2. Calcule la *magnitud* de la asociación, diferencia o riesgo
3. Calcule su grado de significación *(p)* y los intervalos de confianza
4. Si la diferencia no es estadísticamente significativa, calcule el poder estadístico
5. Evalúe la relevancia clínica de sus resultados, tanto si son estadísticamente significativos como si no lo son
6. Considere los posibles errores en su estudio y las explicaciones alternativas a los resultados obtenidos
7. Proponga otros estudios concretos para avanzar más en el área objeto del trabajo

Estos criterios *no* son exhaustivos y se refieren sobre todo al cálculo de la significación estadística (v. explicación en el texto).

Por lo tanto, cuando *p* no es «pequeña» ($p > 0,05$ o $p > 0,01$), puede ocurrir: *a)* que la asociación observada no sea fuerte (o que el efecto sea débil, o que la diferencia hallada tenga escasa magnitud), y *b)* que la asociación o la diferencia sea notable pero que el número de pacientes analizados no sea lo bastante grande como para descartar que el resultado sea casual.

Cuando el grado de significación estadística no es significativo

Los resultados que no tienen significación estadística no constituyen, por supuesto, ninguna desgracia, a pesar de que a menudo se los rotula como «negativos» para destacar que no se encontró un efecto, una diferencia o una asociación. ¿Qué hacer, cómo interpretarlos?

Cuando *p* es «grande» (p. ej., $p > 0,05$), hay que preguntarse en primer lugar: ¿se observa alguna asociación o diferencia? Si la diferencia entre dos porcentajes o entre dos medias se aproxima a cero, o si el riesgo relativo está cerca de 1, entonces se dice que no se observa ninguna asociación o diferencia (o que éstas tienen muy poca magnitud). En ese caso la mejor manera de presentar el resultado es decir: no se observó una asociación (o diferencia, o aumento del riesgo).

Pero a continuación hay que preguntarse: ¿cuál era la potencia o poder estadístico?, es decir, ¿qué posibilidades había de detectar una diferencia si ésta en realidad existía?

Si el poder era bajo (p. ej., inferior al 80 o al 70 %), entonces habría sido mejor perder el tiempo en otra cosa, o invertirlo convenciendo a otros colegas de que aportaran sus pacientes a un estudio multicéntrico. Un mayor número de pacientes habría proporcionado el poder estadístico suficiente para que los resultados fueran interpretables. Pues, en efecto, un principio fundamental en investigación es el siguiente: nunca hay que empezar un estudio cuyos resultados serán, de todos modos, ininterpretables, sea por falta de poder estadístico o por otra causa. Antes de empezar un estudio clínico es necesario calcular el poder estadístico que, con el número de pacientes que se piensa estudiar, se tendrá para detectar determinadas diferencias.

Si el poder estadístico era alto, entonces se puede decir que los datos van en contra de la existencia de una asociación fuerte (o de una diferencia grande). Teóricamente, es posible que un estudio con una muestra de tamaño descomunal encuentre que una diferencia muy pequeña es significativa desde el punto de vista estadístico.

Si se observa una asociación o diferencia, entonces puede decirse que se ha apreciado una asociación de tal magnitud entre tal y cual, pero que, sin embargo, el número de pacientes analizados ha sido insuficiente para poder descartar que la asociación se deba al azar. No se debe decir que no se ha detectado una asociación, puesto que sí se la ha hallado, sino que sólo había un porcentaje x de posibilidades (el poder estadístico) de encontrar una asociación estadísticamente significativa. Y a continuación hay que evaluar la significación clínica de la asociación o diferencia encontrada, dejando siempre bien claro que se trata de un juicio personal subjetivo.

En resumen, pueden darse cuatro situaciones: *a)* la diferencia es clínica y estadísticamente significativa; *b)* la diferencia es clínicamente significativa pero no tiene significación estadística (el número de pacientes estudiados es insuficiente para determinar si la diferencia es real o casual);

c) la diferencia es estadísticamente significativa pero irrelevante desde el punto de vista clínico (el tamaño muestral es adecuado, o incluso demasiado grande, pero se considera que la diferencia no tiene importancia para el cuidado de los enfermos), y *d)* la diferencia no es ni clínica ni estadísticamente significativa (el tamaño de la muestra es suficiente y realmente no existen diferencias o bien la muestra es insuficiente y la diferencia podría ser real o casual).

Lo significativo de lo significativo

Por último, conviene destacar cinco cuestiones de enorme importancia:

1. Aunque *p* sea «pequeña» (p. ej., $p < 0{,}01$), la diferencia hallada puede ser demasiado pequeña para tener importancia clínica.
2. Deben distinguirse claramente las hipótesis formuladas antes del inicio del estudio y puestas a prueba en el transcurso de éste (hipótesis *a priori*) de aquellas que se hayan generado después de haber aplicado ya algunas pruebas estadísticas (hipótesis *a posteriori*). Cualquier protocolo de investigación debe determinar por adelantado las diferencias o efectos que se considerarán como clínicamente significativos, y en especial aquellos que influirán sobre las decisiones clínicas (p. ej., cambiar de tratamiento o utilizar una nueva prueba diagnóstica).
3. Las pruebas de significación estadística no deben suplantar jamás al juicio clínico o sanitario. Éste debe incorporar toda la experiencia acumulada por la comunidad médica, incluyendo una valoración crítica de la bibliografía. Las pruebas estadísticas sólo responden la siguiente pregunta: ¿cuál es la probabilidad de que la diferencia observada se deba al azar?
4. Cualquier resultado, incluso el de mayor trascendencia práctica, puede perder su significación estadística si se divide a los pacientes en subgrupos. Empezar a descomponer el grupo de pacientes estudiado en pequeños subgrupos, para que un hallazgo deje de ser estadísticamente significativo, es un fraude y, por lo tanto, ética y profesionalmente inaceptable. Los subgrupos que se analicen se deben corresponder con las hipótesis formuladas *a priori* y, por consiguiente, deben haber sido definidos de antemano en el protocolo.
5. Por último, no se debe perder de vista que en los ejemplos dados se ha partido de la base de que los datos analizados son válidos, es decir, de que el estudio está razonablemente libre de sesgos. La estadística no puede corregir las insuficiencias del diseño. Un estudio con errores sistemáticos en la selección o en la comparación de los pacientes, o cuyos datos clínicos son erróneos, o que no tiene en cuenta posibles factores de confusión, no mejora mediante *p* significativas.

El diccionario de la lengua española define así a un papanatas: «Hombre simple y crédulo o demasiado cándido y fácil de engañar». La epidemiología clínica y otras disciplinas afines ayudan al investigador a no ser un papanatas al recordarle que la relevancia o significación clínica de un estudio con un grado razonable de validez interna es siempre más importante que la precisión estadística y que muchas otras consideraciones matemáticas.

Bibliografía

Carné X, Porta M. L'evolució dels models de causalitat en Medicina. Gac Sanit (Barc) 1983; 2: 54-57.

Evans AS. Causation and disease. A chronological journey. New York: Plenum, 1993.

Plasència A, Porta Serra M. La calidad de la información clínica (II): significación estadística. Med Clin (Barc) 1988; 90: 122-126.

Porta Serra M. Métodos de investigación clínica: errores, falacias y desafíos. Med Clin (Barc) 1990; 94: 107-115.

Porta Serra M, Álvarez-Dardet C, Bolúmar F, Plasencia A, Velilla E. La calidad de la información clínica (I): validez. Med Clin (Barc) 1987; 89: 741-747.

Porta Serra M, Plasencia A, Sanz F. La calidad de la información clínica (y III): ¿Estadísticamente significativo o clínicamente importante? Med Clin (Barc) 1988; 90: 463-468.

Susser M. Causal thinking in the health sciences. Concepts and strategies in epidemiology. New York: Oxford University Press, 1973.