En busca de la p significativa. Su influencia sobre la credibilidad de lo publicado

In Search of the Significant p. Its Influence on the Credibility of Publications

JORGE THIERER

La publicación de los resultados de un estudio en una revista donde impera la revisión por pares es la coronación de la investigación en cualquier rama de la ciencia. Se asume, obviamente, que los resultados presentados son correctos y que un análisis cuidadoso los valida.

Un fenómeno que llama la atención es que la inmensa mayoría de las publicaciones presentan estudios "exitosos", en los que la hipótesis inicial es refrendada con su demostración. Una revisión de Fanelli refiere que el 84% de los estudios publicados demuestran lo que la introducción presenta como hipótesis inicial. (1) El dato es llamativo. En la misma línea suena preocupante que la cantidad de publicaciones en que los hallazgos pueden ser replicados sea baja, desde la investigación en cáncer (2) hasta la psicología. (3)

En Medicina, como en otras disciplinas, la aproximación hipotético-deductiva es la base de la investigación. A partir de una teoría plausible se genera una hipótesis *a priori*, que será testeada en un experimento con un diseño preciso, al que se considera el mejor para demostrarla. Una prueba estadística determinada (de acuerdo con el tipo de variable que se compara, y el número y la distribución de sus valores) nos informará un valor de p, pero, ¿cuál es el significado de p?

¿QUÉ SIGNIFICA P?

Cualquier resultado de cualquier variable forma parte de una distribución de probabilidades. Es más probable que algunos resultados ocurran con más frecuencia que otros. El ejemplo clásico de distribución de probabilidades es la curva de Gauss, donde los valores más cercanos a la media son los que tienen mayor probabilidad de ocurrir, mientras que los más alejados ocurren con mucha menor frecuencia.

Frente a cualquier comparación entre un valor obtenido y uno de referencia, o entre un valor en un grupo y otro grupo con el que se compara, es posible que en cada situación ambos formen parte de la misma distribución de probabilidades y, por lo tanto, que sea el azar el responsable de que hayamos encontrado valores diferentes.

No es lo mismo el significado de p en una prueba de significación que en una de hipótesis. (4)

En una *prueba de significación* importa el valor que asume p: representa la probabilidad de encontrar valores (o diferencias entre valores) como los hallados o más extremos si la hipótesis nula es cierta. En este tipo de prueba (denominada fisheriana en honor a Fisher, difusor del concepto en un texto clásico de 1925) el valor de p es un continuo y, cuanto más bajo, es menor la probabilidad de encontrar los valores explorados bajo la hipótesis nula. Entonces, por ejemplo, al explorar una diferencia de medias entre dos grupos, si el valor de p hallado es 0,10 diremos que si la hipótesis nula es cierta (si no existe verdaderamente diferencia entre los dos grupos) la probabilidad de encontrar la diferencia hallada es de 10%.

En una *prueba de hipótesis*, que es en general la que domina las reglas aceptadas de la investigación, la valoración de p es binaria: importa, en general, si el valor de p es $<0,05$ o no lo es. Esta forma de valorar p fue difundida por Neyman y Pearson en 1933. Se reúnen datos y se plantea una prueba de hipótesis que considera una hipótesis nula y una alternativa.

Si el azar llevó al hallazgo de valores diferentes dentro de la misma distribución, no hay diferencia significativa entre sí. Esto es lo que formula la hipótesis nula. Frente a la misma se alza la hipótesis alternativa, que afirma que sí hay diferencia entre el valor hallado y el de referencia, o entre el de uno y otro grupo. Para esta hipótesis, entonces, los valores hallados no forman parte de la misma distribución, sino que corresponden a poblaciones diferentes.

¿Cómo decidir cuál de las dos hipótesis es la correcta? Todo gira en relación con la hipótesis nula, y de rechazarla, o no. Una prueba estadística especial para cada caso permite establecer, en la distribución de probabilidades correspondiente a la variable que estamos estudiando, dónde se encuentra la determinación que nos importa respecto de aquella con la que la estamos comparando; y aclara, si existe, o no, una diferencia estadísticamente significativa entre ambas.

Esa prueba nos informa si en la distribución de probabilidades, respecto del valor del comparador el que nos interesa está en el área de no rechazo o de rechazo de la hipótesis nula. Por un criterio convencional se entiende que la hipótesis nula no se puede rechazar si

el valor hallado se encuentra dentro del 95% central de la distribución de probabilidades que tiene como referencia al poblacional, o al del grupo comparador. Entendemos en este caso que la diferencia se debe al azar. Si en cambio, el valor obtenido se encuentra por fuera de ese 95%, si forma parte del 5% de valores extremos, la hipótesis nula es rechazada.

Cuando en una comparación entre dos valores la prueba informa que la probabilidad de que pertenezcan a la misma distribución es menor que 5%, decimos sí, que la diferencia es estadísticamente significativa. Este valor de corte para rechazar la hipótesis nula (5%, 1 en 20, p = 0,05) es expresión de una convención.

Por otra parte, más allá de la prueba, en el plano real se plantea que la hipótesis nula puede ser verdadera o falsa. Y la prueba estadística puede acertar o no, al describir esa realidad. Se plantean entonces cuatro situaciones. (5)

Si hay una diferencia real y el test la detecta al rechazar la hipótesis nula, el resultado del test es un verdadero positivo.

Si la diferencia no es real, y ambos valores corresponden a la misma distribución de probabilidades, pero la prueba estadística rechaza la hipótesis nula planteando que el valor de p es <0,05 estamos en presencia de un resultado falso positivo. A este resultado erróneo lo denominamos error alfa o tipo I. Si la tasa de resultados falsos positivos del test es de 5%, su especificidad, o tasa de verdaderos positivos es del 95%: 1-error alfa. Cuanto menor sea el valor del error alfa, más específico será el test.

Si la diferencia no existe, ambos valores pertenecen a la misma distribución y el test no rechaza la hipótesis nula; en ese caso estamos en presencia de un verdadero negativo.

Si hay una diferencia real, pero el test no la detecta, no alcanza a rechazar la hipótesis nula; hablamos entonces de un resultado falso negativo, y entendemos que se ha producido un error beta o tipo II. Si consideramos 1- error beta obtenemos la potencia o poder del test, que representa su sensibilidad y su capacidad para encontrar una diferencia real. Así como habitualmente se admite un error alfa de 0,05 o 0,01, el error beta que se define *a priori* en una comparación puede ser de 0,10 (si se busca una sensibilidad de 90%) o 0,20 (si se admite una sensibilidad del 80%).

Para la misma cantidad de observaciones los errores alfa y beta se desempeñan en forma contrapuesta. Como con cualquier test diagnóstico, si aumenta la sensibilidad o poder, y por ende, cae el error beta, aumenta el error alfa, es decir, cae la especificidad. Si, por el contrario, cae el error alfa, y se busca ser más específico (eligiendo un valor de p <0,01 por ejemplo) cae el poder del estudio para detectar una diferencia. Seguramente en la fase exploratoria de un estudio se trabaja con mayor sensibilidad (se admite mayor poder y, por ende, más falsos positivos) mientras que a la hora de confirmar una hipótesis *a priori* se privilegia la especificidad.(6)

Como vemos, entonces, la concepción fisheriana entiende el valor de p como un continuo, y la concepción pearsoniana, en forma dicotómica. Si bien existe un consenso cada vez mayor en que esta concepción dicotómica puede llevar a errores graves y le falta flexibilidad para poder sacar conclusiones adecuadas sobre los fenómenos sometidos a estudio (si p = 0,049, aceptamos la diferencia; si p = 0,051, decimos que no podemos hacerlo), lo cierto es que en la valoración de la evidencia y de los escritos presentados para publicación, el valor de 0,05 sigue siendo el criterio universalmente aceptado para decidir si un resultado es positivo o no. Y aunque algunos han propuesto un valor de 0,005, es cierto que el mismo reduce la tasa de falsos positivos, pero no resuelve definitivamente el problema. (7) El empleo de un punto de corte fijo reduce seriamente la posibilidad de reexplorar un resultado: si ya se obtuvo en un estudio un valor de p = 0,04, ¿quién repetirá el experimento para ver si se trata de un falso positivo? Si el valor de p fue 0,08, ¿cuántos iniciarán un nuevo estudio con la idea de que pudo tratarse de un falso negativo?

En los últimos años se han difundido una serie de advertencias sobre conceptos erróneos respecto del significado de p y del error de confiar en un test de hipótesis, (8, 9), así como defensas encendidas de su uso (10), que serán la base de una próxima *Página del Editor*. De cualquier manera, repetimos, el concepto de p <0,05 sigue siendo la piedra angular sobre la que reposa el concepto de estudio con resultado positivo o no, y en la búsqueda de alcanzar el valor deseado de p pueden darse una serie de situaciones que conspiran contra la credibilidad de los hallazgos.

Veamos primero una aseveración que puede ser disruptiva. Hace 15 años Ioannidis afirmó –en un artículo que generó escozor– que la mayoría de los hallazgos de investigación publicados son falsos(11). ¿Una *boutade* o una verdad irrebatible? ¿Cómo justificarla? Supongamos que de 1 000 hipótesis científicas que se plantean el 10% son certeras y el 90% restante, no. Si trabajamos con un valor alfa de 5% (esto es, que se acepta una tasa de falsos positivos cuando la hipótesis del 5% nula es cierta) sobre las 900 hipótesis erróneas se obtendrá un valor de p <0,05, y por ende, se rechazará la hipótesis nula en 45. ¡Cuarenta y cinco veces aceptaremos una hipótesis que no es correcta sobre 900 exploradas! Y si trabajamos con un poder de 80% para detectar verdaderos positivos, de las 100 hipótesis correctas solo 80 serán detectadas. Por ende, de 45 + 80 = 125 hipótesis aceptadas, 45 (el 36%) serán falsas.

Si el poder fuera de solo un 20%, los hallazgos falsamente positivos subirían al 53%; pero si la proporción de hipótesis correctas no fuera del 10% sino del 50%,

con un error alfa de 5% y un error beta de 20%, la proporción de hallazgos falsamente positivos sería del 11%. Significa que la proporción de hallazgos verdadera y falsamente positivos sobre el total de hipótesis planteadas depende de los errores alfa y beta que aceptemos, y de la proporción de hipótesis verdaderamente correctas; es decir, de la proporción de veces en que la hipótesis nula es falsa. Saber que la cantidad de veces que podemos estar enfrentando un falso positivo supera largamente el 5%, a pesar de haber empleado un valor de p <0,05, sin duda obliga a extremar los cuidados para mantener la validez de lo publicado.

MECANISMOS QUE PERMITEN ALCANZAR UN VALOR DE P SIGNIFICATIVO

Múltiples comparaciones

Si exploramos dos grupos y valoramos una serie de características basales buscando diferencias significativas entre ambos, y para cada una de las comparaciones se establece un valor de p <0,05 para rechazar la hipótesis nula, es decir que se acepta hasta un 5% de falsos positivos, el hecho de realizar varias comparaciones aumenta notablemente esa tasa.

Por ejemplo, si llevamos a cabo 10 comparaciones asumiendo en cada una un valor de p <0,05 para definir significación estadística, la probabilidad de tener un resultado positivo simplemente por chance trepa al 40%. Asumir como significativo un valor de p <0,05 en este contexto es claramente erróneo. Por eso, cuando se han llevado a cabo múltiples comparaciones se recomienda trabajar con un valor de p menor al usual.

Supongamos que hicimos 20 comparaciones estableciendo *a priori* un valor de p <0,05; de las 20 comparaciones solo una resultó con p <0,05, pero entendemos que puede ser un falso positivo. Aplicando la corrección de Bonferroni se divide 0,05 por 20, y se obtiene un valor necesario de p = 0,0025 para poder rechazar la hipótesis nula en cada una de las comparaciones. Ello disminuye francamente la tasa de falsos positivos, y sin duda debe aplicarse; pero también es cierto que un valor de p = 0,003 bajo esta corrección se torna no significativo.

Generalmente no llegamos a conocer todos los análisis que se han llevado a cabo. Si sobre 10 modelos predictivos testeados solo uno es adecuado, ese es el que se informa. ¿Cómo saber que hay otros 9 que duermen subrepticiamente el sueño de los justos? ¿Cómo interpretar la significación estadística del hallazgo positivo si no sabemos de cuántas comparaciones emerge y, por lo tanto, cuál es la chance de estar ante un falso positivo?

En este contexto la falacia de la evidencia incompleta, "cosecha de cerezas" o *Cherry Picking*, remite justamente a seleccionar las variables o relaciones que sirven para convalidar una hipótesis y omitir las que no la demuestran. Es una variable del fenómeno de múltiples comparaciones, con una selección sesgada de la evidencia. Se la denomina también falacia de la evidencia incompleta.

No es fácil resolver el problema. (12) Tal vez, parte de la tensión radica en la necesidad de rechazar o no la hipótesis nula, haciendo descartables los hallazgos en que el valor de p fue ≥0,05. Otro punto tiene que ver con diferenciar si la variable que aparece como estadísticamente significativa es independiente o no de las otras exploradas. Si no lo es, claramente se debe llevar a cabo el ajuste. Si lo es, podríamos aceptar el resultado obtenido.

EL PIRATEO DE LA p (*p HACKING*)

Bajo esta denominación se engloban una serie de procedimientos dirigidos a obtener un valor de p significativo, cuando, llevado a cabo el análisis inicial, el mismo no se logró. (13) Se entiende que el valor de p debe estar cercano a 0,05. No se llevarían a cabo estas acciones si el valor obtenido inicialmente fuera por ejemplo, 0,45. Entendemos como un valor inicial que motiva esta conducta uno que oscila entre 0,05 y 0,10, o tal vez, hasta 0,20.

Entre estas acciones podemos distinguir: (4, 14)

a) *Analizar solo un subgrupo de datos*. Solo se justifica si hay una razón muy fuerte para excluir el resto, como por ejemplo ser irrelevantes para la cuestión planteada, o existir dudas sobre la toma de la información. De cualquier manera, este procedimiento es muy dudoso cuando se lleva a cabo *post hoc*, y no había definiciones claras *a priori* en el protocolo.
b) *Excluir outliers* (datos muy alejados de la media de la muestra). Es una práctica aceptable si son datos mal recabados, que están por fuera de los criterios de inclusión iniciales. No es una práctica aceptable si se trata de datos correctos, y no hay razón de fuste para retirarlos, más allá de que su exclusión reduce la variabilidad de la muestra y, al disminuir el error estándar, aumenta el poder estadístico, con lo que disminuye los falsos negativos, pero aumenta los falsos positivos.
c) *Estandarizar* (por ejemplo, ajustar el peso por la altura) y *transformar logarítmicamente* datos de distribución anormal. Se trata de mecanismos que aumentan la capacidad de lograr un valor de p significativo, y desde ya que su uso no es reprochable, pero deben haber sido establecidos *a priori*. Si se recurre a ellos cuando el análisis inicial no fue satisfactorio, aumentan la chance de falso positivo porque se está buscando significación por segunda vez, estableciendo entonces más de una comparación.
d) *Aumentar el tamaño de la muestra*. Este es el consejo que con más frecuencia reciben los autores de un trabajo que no arrojó resultados estadísticamente significativos. Como sabemos, el tamaño de muestra

de un estudio está definido por cuatro factores: el error alfa o tasa de falsos positivos aceptable (valor de p), el error beta o tasa de falsos negativos, y por tanto el poder (que es 1 – error beta), la magnitud del efecto que se está explorando y su desvío estándar que expresa la variabilidad del efecto. Cuanto menor es el error alfa, el error beta y la magnitud del efecto, y mayor la variabilidad, es mayor el tamaño de muestra necesario. Si no se puede demostrar que una diferencia determinada o un efecto son estadísticamente significativos, surge el concepto de que tal vez, eso se debe a que no hubo poder suficiente para hacerlo. Como los valores del efecto y su variabilidad son datos que ya no pueden modificarse con los datos de que se dispone, y se busca un valor de p <0,05, lo único que queda es aumentar el poder por medio del incremento del tamaño muestral.

Se puede calcular entonces cuál es la cantidad de observaciones que deben agregarse para alcanzar un valor de p <0,05 con el efecto o la diferencia encontrados, y su variabilidad. Si bien puede postularse que el procedimiento es correcto, no es menos cierto que no sabemos si el efecto encontrado y su variabilidad son verdaderos estimadores de los efectos en la población y que, por lo tanto, podemos estar haciendo un esfuerzo grande para demostrar algo que no es real; por otra parte, que al volver a buscar significación con mayor número de datos estamos aumentando la tasa de falsos positivos, porque damos a muchas observaciones la posibilidad de ser testeadas en más de una oportunidad.

Entonces, este aumento de la tasa de falsos positivos se da bajo el supuesto de que la hipótesis nula es verdadera. Si la hipótesis nula es falsa, esto es, si verdaderamente existe la diferencia, lo que se ha hecho al aumentar el poder es disminuir la tasa de falsos negativos. ¿Es la hipótesis nula verdadera o falsa? Ese es el problema. ¡No lo sabemos! De allí que, entre los mecanismos puestos en juego para alcanzar el valor de p buscado, este sea probablemente el menos criticado. Sí se sugiere que sería adecuado mencionar el tema en la sección Discusión, informando a los lectores de las dificultades halladas.

La práctica de p *hacking* surge de un análisis flexible, en que los grados de libertad de que dispone el autor para lograr su propósito son muchos. (15) El autor puede, tras el análisis inicial, recurrir a uno o más de uno de los procedimientos previamente descriptos, y resulta muy difícil poder detectar su práctica una vez publicado el trabajo.(16)

HIPOTETIZAR DESPUÉS DE CONOCER LOS RESULTADOS (*HARKING*)

En los últimos 30 años una nueva forma de aproximarse a la investigación ha crecido progresivamente. Es lo que Kerr en su escrito liminal (17) denominó *HARKing* (*Hypothesizing After the Results are Known*, Hipotetizar Después que los Resultados son Conocidos). *HARKing* consiste en presentar en la introducción de un trabajo de investigación, una hipótesis generada *post hoc*, una vez que se analizaron los datos, como si la misma hubiera sido formulada *a priori*, antes de recabarlos. Una tabla de contingencia de 3 filas y dos columnas presenta las 6 situaciones que pueden darse al contrastar la hipótesis antes y después de conocer los datos

La idea subyacente al *HARKing* es que no importa si la hipótesis sustentable tras el análisis ha sido previamente formulada. Lo que importa es que sea plausible. Bajo este supuesto, entre las hipótesis a, c y e puede ser elegida y planteada en la introducción, como si lo hubiera sido *a priori*, la más plausible. De la combinación de las condiciones pre y post estudio, surgen para Kerr una serie de variantes, de las cuales entendemos como las más factibles: suprimir las hipótesis plausibles que no se han podido demostrar (si se habían planteado *a priori* a y b, mencionar en la introducción solo la a, generando la ilusión de que solo se pensó lo que efectivamente se comprobó) o extender esa condición a las no plausibles (presentar en la introducción las hipótesis a y c) e incluso, a las que ni siquiera se anticiparon. La condición común a todas estas variantes es que son los resultados los que definen qué hipótesis se presentan en la introducción, y aparecen sostenidas por la evidencia.

¿Cuándo podemos sospechar *HARKing*? Un criterio es que la hipótesis presentada como formulada *a priori* no surge claramente de la teoría o paradigma en el que se inscribe el estudio. Esto podría suceder porque faltan variables esenciales que hubieran sido consideradas para demostrar la hipótesis, de ser cierta su condición *a priori;* o porque surge la sospecha de que los métodos empleados no son los que se hubieran pensado en esa circunstancia. Lo cierto es que algunas formas de *HARKing* que mencionamos son tan empleadas como la aproximación hipotético-deductiva. De hecho, dejar de lado –si los datos así lo sugieren– una hipótesis, aunque haya sido considerada plausible *a priori*, y centrarse en hipótesis que no fueron consideradas, pero aparecen plausibles, es una conducta que muchos textos recomiendan.

¿Por qué se recurre al *HARKing*? Se pueden mencionar muchas razones. En principio, se entiende que

	Hipótesis tras conocer los resultados	
Hipótesis antes de realizar el estudio	Plausible	No plausible
Anticipada y plausible	a	b
Anticipada y no plausible	c	d
No anticipada	e	f

siempre debe haber una teoría y una hipótesis que hayan precedido al estudio. Y aun cuando Popper haya señalado que lo ideal era que una teoría fuera rebatida, lo cierto es que el ideal es que se formule una teoría que el estudio confirma; es lo que los lectores esperan. Y lo que los investigadores creen: en una encuesta el 89% entendió que una prueba que confirma da información, pero solo el 39% pensó lo mismo de una prueba que no confirma. La comunicación de resultados confirmatorios satisface la necesidad de los investigadores, de los lectores y de las autoridades de las revistas de conformar un relato que tranquiliza y da una falsa ilusión de progreso científico. Si el resultado es positivo, es más factible que se lo publique; mayor número de publicaciones asegura más acceso a subsidios y a éxito en el mundo académico. La publicación premia la novedad y la idea de la hipótesis original convalidada por un estudio con "final feliz". En virtud de este sesgo de confirmación retrospectiva se llega a creer que el hallazgo inesperado era algo que en realidad se sabía desde siempre. Que la hipótesis que se presenta no haya sido formulada *a priori* se transforma muchas veces en un tema menor para la mayor parte de los interesados. Y hasta hay editores, generalmente de las revistas que cobran a los autores por cada publicación, que sugieren llevar a cabo este procedimiento como forma de facilitarla.

Esta preferencia por hipótesis novedosas confirmadas por los hechos se enraíza con el hecho de que la hipótesis nula tiene mala prensa. Los estudios que señalen que no hay diferencia, que es lo mismo lo nuevo que lo ya sabido, que una intervención no modifica la evolución conocida, que un método diagnóstico recientemente desarrollado no agrega a los recursos tradicionales, no son estudios que entusiasmen al lector, imbuido por la creencia en el perpetuo progreso.

La hipótesis nula "debe" ser rebatida a cualquier precio. No está de más señalar que, cuando se realizan simultáneamente varias pruebas de significación, todas con un valor de $p < 0,05$, la tasa de falsos positivos, de error tipo I, crece notablemente. En ese sentido la práctica de *HARKing* es una causa muy frecuente de ese error. Por ende, y desconociendo que hay alta chance de falso positivo, el mismo se convierte en evidencia aceptada.

Frente a la hipótesis generada prospectivamente (a la que se llama "predicción") se alza la generada *post hoc* (a la que Horwich denominó "acomodación"). Podría argumentarse que, si ambas fueran examinadas de igual forma, y contra el mismo cuerpo de evidencia, no habría diferencias de credibilidad entre sí. Pero lo más frecuente es que cuando se genera una hipótesis *post hoc*, la interpretación de la evidencia esté sesgada a favor de los hallazgos más recientes. Se genera una narrativa que convalida hallazgos que no fueron sospechados al inicio del estudio. Una diferencia básica entre el *HARKing* y la práctica de p *hacking* es que en el primer caso la hipótesis inicial no es la que finalmente se presenta; en el caso de p *hacking* la hipótesis se conserva, pero varía el análisis previsto inicialmente. (18)

La práctica de *HARKing* es claramente no tan grave como la falsificación de datos o el plagio. De hecho, no hay códigos de ética en investigación que la condenen explícitamente. Sin embargo, es claro que sus consecuencias pueden ser gravosas: hay información que se pierde, se es más laxo en la aplicación de métodos estadísticos, no se es honesto en la comunicación de los resultados. El hallazgo sustentado en un valor de $p < 0,05$ puede llevar a defender una línea de investigación que deja información relevante en la periferia.

Como se ha logrado el objetivo de demostrar significación estadística, deja de importar si se cree realmente en lo que se presenta, fomentando una actitud centrada en el resultado que deja fuera del campo la búsqueda de la verdad. Al pretender que "se encontró lo que se buscaba" se deja de lado la posibilidad de reconocer hallazgos por serendipia, (19) se cae en la adopción de teorías muy estrechas que justifican exactamente los hallazgos o, por el contrario, ante la incapacidad de poder explicarlos certeramente se edifican teorías difusas que pueden explicar casi todo en general y nada en particular; y se abandona la posibilidad de generar hipótesis alternativas. A diferencia de la falsificación o el plagio, no es sencillo detectar o demostrar la práctica de *HARKing*, y la respuesta no puede ser la creencia de que todo lo que leemos proviene de ella.

CONSIDERACIONES FINALES

Hemos presentado solo algunas de las situaciones que pueden llevar a aumentar la publicación de resultados falsamente positivos. Dejamos exprofeso de lado las situaciones claramente desdorosas, como la falsificación y el plagio. Elegimos centrarnos en aquellas en las que, en forma inadvertida, o por entenderse que se está haciendo buena ciencia, y porque se interpreta que las acciones llevadas a cabo no conspiran contra la credibilidad de lo que se presenta, se toman decisiones que liman los supuestos del método científico.

La proporción de investigadores que las practican es alta. Por ejemplo, en una encuesta llevada a cabo con 494 investigadores en ecología y 313 en biología evolutiva, no reportar variables que fueron analizadas y que no lograron significación estadística fue reconocido por más del 60%; el agregado de nuevas observaciones para aumentar el poder, por más del 40%; y haber practicado *HARKing*, por un 50%. (14)

¿Por qué son prácticas difundidas? Hay sin duda un componente individual: no todos los autores las practican. Hay, por otra parte, investigadores muy honestos que no ven en su ejercicio una actitud realmente criticable. Y hay quienes entienden que forma parte

normal de la práctica de la investigación. Numerosos sesgos cognitivos (de observación, de confirmación, de comprensión retrospectiva, de creencia) (20) pueden explicar una tendencia inconsciente a privilegiar en el análisis estadístico o en la redacción de un escrito, la información que confirma las propias creencias dejando de lado lo que las pone en entredicho, y también a hilar el discurso que lleva implícitamente a la publicación.

Pero, más allá de lo personal, podemos detenernos –y esto parece incluso más decisivo– en el sistema de ideas imperante. Si solo se premia el hallazgo de resultados positivos, si priva la religión de la $p < 0,05$, si los que más frecuentemente llegan a ese resultado son los que acceden más llanamente a la publicación, y si los estudios dedicados a replicar hallazgos ya publicados son desalentados; si "publicar o perecer" es el primer mandamiento, pero la publicación solo premia a la p significativa, las prácticas que describimos seguramente no desaparecerán.

Se trata de una actitud diferente de los editores, capaces de aceptar comunicaciones iniciales de investigación que reconozca ser exploratoria; la capacidad de detectar *HARKing* y reclamar entonces de los autores que lo reconozcan; la aceptación de estudios que repliquen hallazgos ya publicados, son todas acciones que pueden llevar a hacer más creíble la literatura médica. Más allá de este punto, y en forma general, si se aceptara que el hallazgo puede haber surgido de la exploración, y se explicaran adecuadamente los supuestos bajo los cuales se llevó a cabo, y las líneas de investigación que se abren, habría suficiente espacio académico para publicar y contribuir al progreso del conocimiento, aunque la p fuera $\geq 0,05$. (Aunque, y desde ya, de cualquier manera, no podemos dejar de preguntarnos entonces hasta qué valor de p admitiríamos para publicar un hallazgo. Tal vez ello debiera depender de la pertinencia de la hipótesis, la importancia clínica del hallazgo, la seguridad de que se trata de una línea de investigación que continúa).

Entonces: por un lado, una actitud más favorable a la investigación de calidad, aunque no alcance a $p < 0,05$. ¿Y por el otro? Una actitud firme para que cuando se explicita $p < 0,05$, se pueda estar seguro de que el hallazgo es cierto, y no un falso positivo. En este sentido, entre las soluciones disponibles se cuenta fomentar la publicación en revistas que aseguren la revisión por pares. El juicio de revisores reconocidos como expertos en su campo de conocimiento y ciegos al nombre de los autores, puede contribuir a detectar anomalías en la presentación de los datos y asegurar la transparencia de lo publicado.

Promover por parte de los autores el reconocer cuándo han recurrido a agregar observaciones, eliminar variables, o cuándo han cambiado su hipótesis guiados por hallazgos inesperados hará más creíble lo publicado. Poder comunicar lo que sucede habitualmente en la investigación, que está lejos de ser un sendero sin obstáculos, puede iluminar la propia trayectoria.

Incluso hay firmes propuestas de poner a disposición el material que sirvió de base a la publicación para que pueda ser revisado aun *a posteriori* de la misma. (21) Pre registrar el protocolo de investigación, señalar claramente los criterios de inclusión y exclusión, los puntos finales y cómo se trabajará con las diferentes variables es, sin duda, una restricción muy fuerte a la tentación también inconsciente de manipular los datos *a posteriori*, de cambiar la hipótesis inicial o los puntos finales. (22)

En la misma línea se inscribe una propuesta más radical aún, que pocas revistas han puesto en práctica, y que consiste en entregar la base de datos sobre la que se hizo el análisis, si se pretende la publicación. (23)

En resumen, la capacidad de Congresos y Publicaciones de dar lugar a ideas originales, aunque estén en proceso de gestación, y al mismo tiempo elevar la vara para que cuando se afirma un hallazgo se pueda sostener lo dicho con hechos, son caminos que deberán transitarse. En la Revista Argentina de Cardiología hacemos todo lo posible para conseguirlo.

BIBLIOGRAFÍA

1. Fanelli D. "Positive" results increase down the Hierarchy of the Sciences. PLoS One 2010;5:e10068. https://doi.org/10.1371/journal.pone.0010068

2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011;10:712. https://doi.org/10.1371/journal.pone.0010068

3. Open Science C. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science. 2015;349:aac4716. https://doi.org/10.1126/science.aac4716

4. Lew MJ. A Reckless Guide to P-values: Local Evidence, Global Errors. Handb Exp Pharmacol. 2020;257:223-56. https://doi.org/10.1007/164_2019_286

5. Akobeng AK. Understanding type I and type II errors, statistical power and sample size. Acta Paediatr 2016;105:605-9. https://doi.org/10.1111/apa.13384

6. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings - a practical guide. Biol Rev Camb Philos Soc 2017;92:1941-68. https://doi.org/10.1111/brv.12315

7. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA 2018;319:1429-30. https://doi.org/10.1001/jama.2018.1536

8. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol 2008;45:135-40. https://doi.org/10.1053/j.seminhematol.2008.04.003

9. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567:305-7. https://doi.org/10.1038/d41586-019-00857-9

10. Ioannidis JPA. Retiring statistical significance would give bias a free pass. Nature 2019;567:461. https://doi.org/10.1038/d41586-019-00969-2

11. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124. https://doi.org/10.1371/journal.pmed.0020124

12. Goodman SN. Multiple comparisons, explained. Am J Epidemiol 1998;147:807-12; discussion 15. https://doi.org/10.1093/oxfordjournals.aje.a009531

13. Wicherts JM, Veldkamp CL, Augusteijn HE, Bakker M, van Aert RC, van Assen MA. Degrees of Freedom in Planning, Running, Analyzing,

and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Front Psychol 2016;7:1832. https://doi.org/10.3389/fpsyg.2016.01832

14. Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. Questionable research practices in ecology and evolution. PLoS One. 2018;13(7):e0200303. https://doi.org/10.1371/journal.pone.0200303

15. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011;22:1359-66.https://doi.org/10.1177/0956797611417632

16. Ulrich R, Miller J. p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). J Exp Psychol Gen 2015;144:1137-45. https://doi.org/10.1037/xge0000086

17. Kerr NL. HARKing: hypothesizing after the results are known. Pers Soc Psychol Rev 1998;2:196-217.https://doi.org/10.1207/s15327957pspr0203_4

18. Wicherts JM. The Weak Spots in Contemporary Science (and How to Fix Them). Animals (Basel) 2017;7(12).https://doi.org/10.3390/ani7120090

19. Pepys MB. Science and serendipity. Clin Med (Lond) 2007;7:562-78. https://doi.org/10.7861/clinmedicine.7-6-562

20. Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. Med Decis Making 2015;35:539-57. https://doi.org/10.1177/0272989X14547740

21. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. SCIENTIFIC STANDARDS. Promoting an open research culture. Science 2015;348:1422-5. https://doi.org/10.1126/science.aab2374

22. Yamada Y. How to Crack Pre-registration: Toward Transparent and Open Science. Front Psychol 2018;9:1831. https://doi.org/10.3389/fpsyg.2018.01831

23. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis. Mol Brain 2020;13:24. https://doi.org/10.1186/s13041-020-0552-2