POR QUÉ LA MAYORÍA DE LOS RESULTADOS DE INVESTIGACIÓN PUBLICADOS SON FALSOS*

John P. Ioannidis[a]

Los hallazgos de investigación publicados a veces son refutados por evidencia posterior, con la confusión y la decepción consiguientes. La refutación y la controversia se ven en la gama de diseños de investigación, desde los ensayos clínicos y los estudios epidemiológicos tradicionales (Ioannidis, Haidich y Lau, 2001; Lawlor, Smith et al., 2004; Vandenbroucke, 2004) hasta la investigación molecular más moderna (Michiels, Koscielny y Hill, 2005; Ioannidis, Ntzani, Trikalinos et al., 2001). Es cada vez más preocupante que en la investigación moderna la mayoría de los reclamos de investigación publicados sean hallazgos falsos (Colhoun, McKeigue y Smith, 2003; Ioannidis, 2003 y 2005). Pero esto no debería ser sorprendente. Se puede probar que la mayoría de los hallazgos investigación reclamados son falsos. Aquí examino los factores clave que influyen en este problema y algunos corolarios relacionados.

* DOI: https://doi.org/10.18601/01245996.v20n39.13. Versión original Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med 2*(8): e124. Traducción de Alberto Supelano. Se publica bajo los términos de Creative Commons Attribution License. Recepción: 23-08-2017, aceptación: 11-05-2018. Sugerencia de citación: Ioannidis, J. P. (2018). Por qué la mayoría de los resultados de investigación publicados son falsos. *Revista de Economía Institucional*, *20*(39), 297-313.

[a] Departamento de Higiene y Epidemiología, Escuela de Medicina, Universidad Iánina, Iánina, Grecia, y Escuela de Medicina de Tufts University, [jioannid@cc.uoi.gr].

MODELACIÓN DEL MARCO DE RESULTADOS FALSOS POSITIVOS

Algunos metodólogos han señalado (Sterne y Smith, 2001; Wacholder, Chanock, Garcia-C. et al., 2004; Risch, 2000) que la alta tasa de no replicación (falta de confirmación) de los descubrimientos es consecuencia de la estrategia conveniente pero infundada de pretender resultados concluyentes solamente con base en un estudio único valorado por la significancia estadística formal, comúnmente para un valor p menor de 0,05. La investigación no es representada y sintetizada apropiadamente por los valores p, pero, infortunadamente, es generalizada la noción de que los artículos de investigación médica se deben interpretar únicamente con base en los valores p. Los resultados de investigación se definen aquí como una relación que alcance significación formal, por ejemplo, intervenciones efectivas, predictores informativos, factores de riesgo o asociaciones. La investigación "negativa" también es muy útil. "Negativa" es en realidad un nombre inapropiado, y la mala interpretación está muy extendida. Sin embargo, aquí me centraré en las relaciones que los investigadores pretenden que existen y no en los hallazgos nulos.

Como se ha demostrado, la probabilidad de que un resultado de investigación sea verdadero depende de la probabilidad previa de que sea verdadero (antes de hacer el estudio), del poder estadístico del estudio y del nivel de significancia estadística (Wacholder, Chanock, Garcia-C. et al., 2004; Risch, 2000). Consideremos una tabla 2 x 2 en la que los resultados de investigación se comparan con el patrón oro de las relaciones verdaderas en un campo científico. En un campo de investigación se pueden formular hipótesis verdaderas y falsas sobre la presencia de relaciones. Sea R la proporción entre el número de "relaciones verdaderas" y "no relaciones" de aquellas que se prueban en ese campo. R es característica del campo y puede variar mucho dependiendo de que el campo se centre en relaciones probables o busque solo una o algunas relaciones verdaderas entre miles y millones de hipótesis que se pueden formular. Consideremos también, por simplicidad de cálculo, campos circunscritos en los que solo existe una relación verdadera (entre las muchas hipótesis que se pueden formular) o hay un poder similar para encontrar alguna de las varias relaciones verdaderas existentes. La probabilidad previa al estudio de que una relación sea verdadera es $R/(R + 1)$. La probabilidad de que un estudio encuentre que una relación verdadera refleja el poder $1–\beta$ (uno menos la tasa de error Tipo II). La probabilidad de pretender una

relación cuando realmente no existe ninguna refleja la tasa de error Tipo I, α. Suponiendo que en el campo se están probando c relaciones, los valores esperados de la tabla 2 x 2 se presentan en el cuadro 1. Después de que se pretende un resultado de investigación basado en el logro de significancia estadística formal, la probabilidad posterior al estudio de que sea verdadero es el valor predictivo positivo (VPP).

El VPP es también la probabilidad complementaria de lo que Wacholder et al. llaman probabilidad de informe falso positivo (Wacholder, Chanock, Garcia-C. et al., 2004). De acuerdo con la tabla 2 x 2, se obtiene $\text{VPP} = (1 - \beta)R/(R - \beta R + \alpha)$. Así, es más probable que un resultado de investigación sea más verdadero que falso si $(1 - \beta)R > \alpha$. Puesto que la gran mayoría de los investigadores suele depender de $\alpha = 0{,}05$, esto significa que es más probable que un resultado de investigación sea más verdadero que falso si $(1 - \beta)R > 0{,}05$.

Cuadro 1
Resultados de investigación y relaciones verdaderas

Resultado de investigación	Relación verdadera		
	Sí	No	Total
Sí	$c(1 - \beta)R/(R + 1)$	$c\alpha/(R + 1)$	$c(R + \alpha - \beta R)/(R + 1)$
No	$c\beta R/(R + 1)$	$c(1 - \alpha)/(R + 1)$	$c(1 - \alpha + \beta R/(R + 1)$
Total	$cR/(R + 1)$	$c/(R + 1)$	c

DOI: 10.1371/journal.pmed.0020124.t001

Lo que es menos reconocido es que los sesgos y el grado de comprobación repetida independiente, por distintos equipos de investigadores del planeta, pueden distorsionar aún más esta imagen y pueden llevar a probabilidades aún más pequeñas que los resultados de investigación sean verdaderos. Intentaremos modelar estos dos factores en el contexto de tablas 2 x 2 similares.

SESGOS

Primero, definamos el sesgo como la combinación de diversos factores de diseño, datos, análisis y presentación que tienden a producir resultados de investigación que no se deberían producir. Sea u la proporción de análisis probados que no habrían sido "resultados de investigación", pero que terminan presentados y reportados como tales debido a sesgos. El sesgo no se debe confundir con la variabilidad al azar que ocasiona que algunos hallazgos sean falsos por casualidad aunque el diseño del estudio, los datos, el análisis y la presentación sean perfectos. El sesgo puede implicar manipulación en el análisis o en el reporte de hallazgos. El reporte selectivo o distorsionado es una forma típica

de ese sesgo. Podemos suponer que u no depende de que exista o no una relación verdadera. Este no es un supuesto irrazonable, puesto que por lo común es imposible saber qué relaciones son verdaderas. En presencia de sesgo (cuadro 2), se tiene VPP = $([1 - \beta]R + u\beta R)/(R + \alpha - \beta R) + u - u\alpha + u\beta R)$, y VPP disminuye cuando aumenta u, salvo que $1 - \beta \leq \alpha$, es decir, $1 - \beta \leq 0{,}05$ para la mayoría de las situaciones. Así, cuando aumenta el sesgo, las posibilidades de que un resultado de investigación sea verdadero disminuyen considerablemente. Esto se muestra en la gráfica 1 para diferentes niveles de poder y diferentes probabilidades pre-estudio.

A la inversa, los hallazgos verdaderos de la investigación pueden ser anulados ocasionalmente debido al sesgo inverso. Por ejemplo, con grandes errores de medición las relaciones se pierden en el ruido (Kelsey, Whittemore et al., 1996), o los investigadores usan datos de modo ineficiente o no advierten relaciones estadísticamente significativas, o puede haber conflictos de interés que tienden a "enterrar" hallazgos significativos (Topol, 2004). No hay buena evidencia empírica a gran escala de la frecuencia con la que puede ocurrir ese sesgo inverso en los distintos campos de investigación. Pero quizá sea justo decir que el sesgo inverso no es tan común. Además, es probable que los errores de medición y de uso ineficiente de datos se estén convirtiendo en problemas menos frecuentes, porque el error de medición ha disminuido con los avances tecnológicos en la era molecular y los investigadores son cada vez más sofisticados acerca de sus datos. No obstante, el sesgo inverso se puede modelar de la misma manera que el sesgo anterior. Además, el sesgo inverso no se debe confundir con la variabilidad aleatoria que puede llevar a ignorar una relación verdadera por azar.

COMPROBACIÓN POR VARIOS EQUIPOS INDEPENDIENTES

Varios equipos independientes pueden abordar el mismo conjunto de preguntas de investigación. Cuando los esfuerzos de investigación se globalizan, la regla práctica es que varios equipos de investigación, a menudo decenas de ellos, puedan indagar las mismas preguntas o preguntas similares. Infortunadamente, en algunas áreas, la mentalidad predominante ha sido centrarse en descubrimientos aislados de equipos individuales e interpretar los experimentos aisladamente. En un número creciente de preguntas hay al menos un estudio que reivindica un hallazgo, y este recibe atención unilateral. La probabilidad de que al menos un estudio, entre varios sobre la misma pregunta, reivindique

un hallazgo estadísticamente significativo es fácil de estimar. Para n estudios independientes de igual poder, la tabla 2 x 2 se muestra en el cuadro 3: VPP=$R(1-\beta^n)/(R+1-[1-\alpha]^n-R\beta^n)$ (sin considerar el sesgo). Con un creciente número de estudios independientes, VPP tiende a disminuir, salvo que $1-\beta<\alpha$, es decir, comúnmente $1-\beta<0{,}05$. Esto se muestra para diferentes niveles de poder y diferentes probabilidades pre-estudio en la gráfica 2. Para n estudios de poder diferente, el término β^n se remplaza por el producto de los términos β_i para $i=1$, pero las inferencias son similares.

COROLARIOS

En el recuadro 1 se muestra un caso concreto. Con base en las consideraciones anteriores se pueden deducir varios corolarios interesantes sobre la probabilidad de que un hallazgo de investigación sea verdadero.

Corolario 1: *Cuanto más pequeños son los estudios realizados en un campo científico, menor es la probabilidad de que los resultados de investigación sean verdaderos.* Un tamaño de muestra pequeño significa menor poder y, para todas las funciones anteriores, el VPP de un hallazgo de investigación verdadero disminuye cuando el poder disminuye hacia $1-\beta=0{,}05$. Por tanto, si los demás factores son iguales, es más probable que los resultados de investigación sean verdaderos en campos científicos que realizan grandes estudios, como las pruebas controladas aleatorias en cardiología (varios miles de sujetos escogidos al azar) (Yusuf, Collins y Peto, 1984), que en campos científicos con estudios pequeños, como la mayoría de la investigación de predictores moleculares (tamaños de muestra 100 veces más pequeños) (Altman y Royston, 2000).

Corolario 2: *Cuanto más pequeño son los tamaños de efecto en un campo científico menor es la probabilidad de que los resultados de investigación sean verdaderos*. El poder también está relacionado con el tamaño de efecto. Por tanto, es más probable que los resultados de investigación sean verdaderos en campos científicos con grandes efectos, como el impacto del tabaquismo en el cáncer o en enfermedades cardiovasculares (riesgos relativos de 3 a 20), que en campos científicos donde los efectos postulados son pequeños, como los factores genéticos de riesgo de enfermedades multigenéticas (riesgos relativos de 1,1 a 1,5) (Ioannidis, 2003). La epidemiología moderna está cada vez más obligada a elegir tamaños de efecto más pequeños (Taubes, 1995). En consecuencia, se espera que disminuya la proporción de resultados de

investigación verdaderos. En la misma línea de pensamiento, si los tamaños de efecto verdaderos son muy pequeños en un campo científico, es probable que este campo esté plagado de pretensiones falsas positivas casi ubicuas. Por ejemplo, si la mayoría de los determinantes genéticos o nutricionales verdaderos de enfermedades complejas confieren riesgos relativos menores de 1,05, la epidemiología genética o nutricional sería en gran parte un esfuerzo utópico.

Cuadro 2
Hallazgos de investigación y relaciones verdaderas en presencia de sesgo

Resultado de investigación	Relación verdadera		
	Sí	No	Total
Sí	$(c[1-\beta]R+uc\beta R)(R+1)$	$c\alpha+uc(1-\alpha)/(R+1)$	$c(R+\alpha-\beta R+u-u\alpha+u\beta R)/(R+1)$
No	$(1-u)c\beta R/(R+1)$	$(1-u)c(1-\alpha)/(R+1)$	$(1-u)(1-\alpha+\beta R/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t002.

Corolario 3: Cuanto mayor es el número y menor la selección de relaciones comprobadas en un campo científico menor es la probabilidad de que los resultados de investigación sean verdaderos. Como se mostró antes, la probabilidad posterior al estudio de que un hallazgo sea verdadero (VPP) depende mucho de las probabilidades pre-estudio (*R*). Por tanto, es más probable que los resultados de investigación sean verdaderos en diseños confirmatorios, como las grandes pruebas aleatorias controladas fase III o sus meta-análisis, que en experimentos que generan hipótesis. Los campos que se consideran muy informativos y creativos dada la riqueza de la información recopilada y comprobada, como los micro arreglos y otras investigaciones orientadas al descubrimiento de alto rendimiento (Michiels, Koscielny y Hill, 2005; Ioannidis, 2005; Golub, Slonim et al., 1999), deberían tener VPP sumamente bajos.

Corolario 4: *Cuanto mayor es la flexibilidad de los diseños, las definiciones, los resultados y los modos analíticos en un campo científico menor es la probabilidad de que los resultados de investigación sean verdaderos*. La flexibilidad aumenta el potencial para transformar lo que serían resultados "negativos" en resultados "positivos", es decir, el sesgo *u*. En algunos diseños de investigación, por ejemplo, en pruebas aleatorias controladas (Moher, Schulz y Altman, 2001; Ioannidis, Evans et al., 2004; ICHE9, 1999) o meta-análisis (Moher, Cook, Eastwood et al., 1999; Stroup, Berlin, Morton et al., 2000), se han hecho esfuerzos para estandarizar su conducta y sus informes. Es probable que la adhesión a estándares comunes aumente la proporción de hallazgos verdaderos. Lo mismo se aplica a los resultados. Los hallazgos verda-

deros pueden ser más comunes cuando los resultados son inequívocos y aceptados universalmente (p. ej., la muerte) que cuando se diseñan resultados muy diversos (p. ej., escalas de resultados de esquizofrenia) (Marshall, Lockwood, Bradley et al., 2000). De mamera similar, los campos que utilizan métodos analíticos estereotipados comúnmente acordados (p. ej., diagramas Kaplan-Meier y pruebas de comparación de sobrevivencia) (Altman y Goodman, 1994) pueden arrojar una mayor proporción de hallazgos verdaderos que campos donde los métodos analíticos aún están en experimentación (p. ej., métodos de inteligencia artificial) y solo se reportan los "mejores" resultados. No obstante, incluso en los diseños de investigación más estrictos, el sesgo parece ser un problema importante. Por ejemplo, hay fuerte evidencia de que el reporte selectivo de resultados, con la manipulación de los resultados y los análisis reportados, es un problema común incluso en pruebas aleatorias (Chan, Hrobjartsson, Haahr et al., 2004). La simple abolición de la publicación selectiva no haría desaparecer este problema.

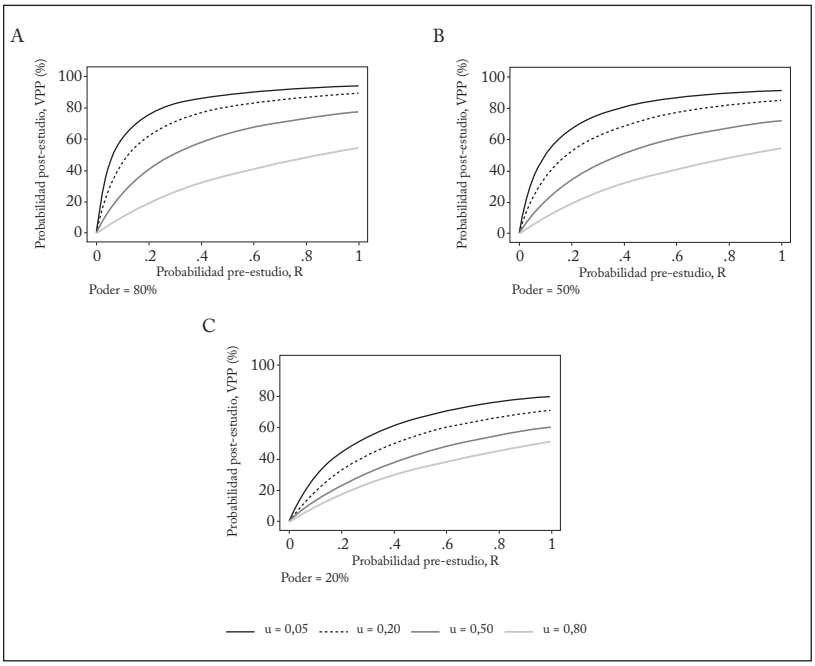
Cuadro 3
Hallazgos de investigación y relaciones verdaderas en presencia de estudios múltiples

Resultado de investigación	Relación verdadera		
	Sí	No	Total
Sí	$c(1-\beta^n)/(R+1)$	$c(1-[1-\alpha]^n/(R+1)$	$c(R+1-[1-\alpha]^n-R\beta^n)/(R+1)$
No	$cR\beta^n/(R+1)$	$c(1-\alpha)^n/(R+1)$	$c([1-\alpha]^n + R\beta^n/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t003.

Corolario 5: *Cuanto mayores son los intereses y prejuicios financieros y de otro tipo en un campo científico menor es la probabilidad de que los resultados de la investigación sean verdaderos*. Los conflictos de interés y los prejuicios pueden aumentar el sesgo u. Los conflictos de interés son muy comunes en la investigación biomédica (Krimsky, Rothenberg, Stott y Kyle, 1998), y suelen ser reportados en forma inadecuada y poco frecuente (ibíd.; Papanikolaou, Baltogianni, Contopoulos-I. et al., 2001). Los prejuicios no necesariamente tienen orígenes financieros. Los científicos de un campo dado pueden tener prejuicios debido simplemente a su creencia en una teoría científica o al compromiso con sus propios resultados. Muchos estudios en apariencia independientes, realizados en las universidades, se pueden llevar a cabo únicamente para dar a los médicos e investigadores calificaciones para su promoción o su titularidad. Tales conflictos no financieros también pueden llevar a reportar resultados e interpretaciones distorsionados. Los

Gráfica 1
VPP (probabilidad de que un hallazgo de investigación sea verdadero)
Como función de las probabilidades pre-estudio de diferentes niveles de sesgo, *u*

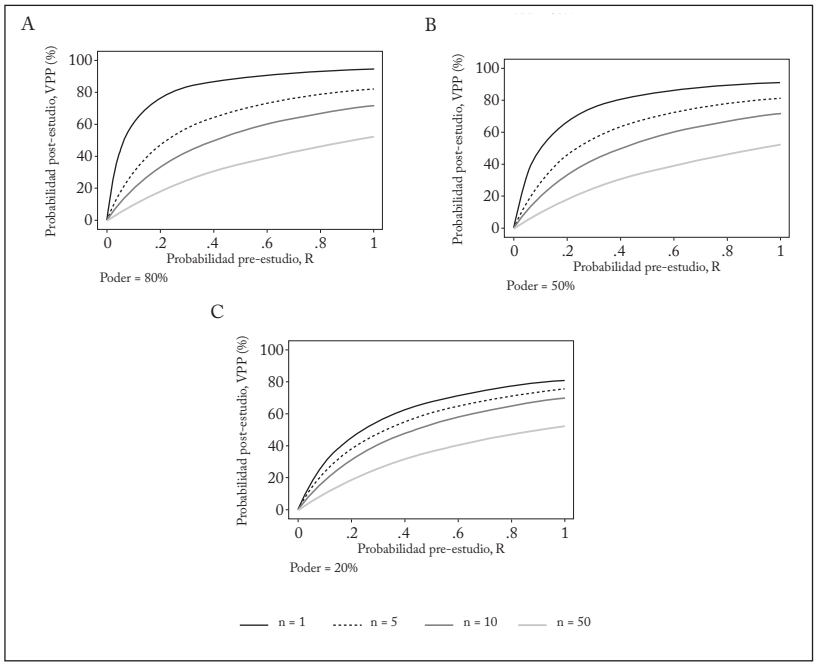


Los paneles corresponden a un poder de 0,20, 0,50 y 0,80.

investigadores prestigiosos pueden suprimir la aparición y la difusión de resultados que refuten sus hallazgos mediante el proceso de revisión por pares. Así perpetúan falsos en su campo de investigación. La evidencia empírica sobre la opinión de los expertos muestra, además, que es poco confiable (Antman, Lau et al., 1992).

Corolario 6: *Cuanto más caliente es un campo científico (que involucra más equipos científicos) menor es la probabilidad de que los resultados de investigación sean verdaderos*. Este corolario aparentemente paradójico se debe a que, como ya se indicó, el VPP de hallazgos aislados disminuye cuando muchos equipos de investigadores participan en el mismo campo. Esto puede explicar por qué a veces vemos gran excitación seguida de profundas desilusiones en campos que atraen mucha atención. Cuando numerosos equipos trabajan en el mismo campo y se producen datos experimentales masivos, la oportunidad es esencial para vencer a los competidores. Así, cada equipo puede dar prioridad a la búsqueda y difusión de sus resultados "positivos" más

Gráfica 2
VPP (probabilidad de que un hallazgo de investigación sea verdadero)
Como función de las probabilidades pre-estudio de diferentes estudios realizados, *n*



Los paneles corresponden a un poder de 0,20, 0,50 y 0,80.

impresionantes. Puede ser atractivo difundir resultados "negativos" solo si otro equipo ha encontrado una asociación "positiva" sobre la misma pregunta. En ese caso, puede ser atractivo refutar una pretensión publicada en una revista prestigiosa. Se acuñó la expresión "fenómeno Proteo" para describir pretensiones de investigación extremas rápidamente alternantes y refutaciones del todo opuestas (Ioannidis y Trikalinos, 2005). La evidencia empírica sugiere que esta secuencia de extremos opuestos es muy común en genética molecular (ibíd.).

Estos corolarios consideran cada factor por separado, pero a menudo se influyen entre sí. Por ejemplo, puede ser más probable que los investigadores que trabajan en campos donde los tamaños reales de efecto se consideran pequeños sean más propensos a realizar grandes estudios que los investigadores que trabajan en campos donde los tamaños reales de efecto se consideran grandes. O en un campo científico caliente pueden predominar perjuicios

que socavan aún más el valor predictivo de sus resultados. Las partes interesadas muy prejuiciadas incluso pueden incluso crear barreras que aborten los esfuerzos para obtener y difundir resultados opuestos. A la inversa, el hecho de que un campo sea caliente o tenga fuertes intereses creados a veces puede promover estudios más grandes y mejores estándares de investigación, lo que mejora

Recuadro 1
Un ejemplo: ciencia con bajas probabilidades pre-estudio

Supongamos que un equipo de investigadores realiza un estudio completo de asociación del genoma para probar si algunos de los 100.000 polimorfismos génicos están asociados con la propensión a la esquizofrenia. Con base en lo que sabemos del grado de heredabilidad de la enfermedad, es razonable esperar que alrededor de diez polimorfismos génicos de los que se prueban estén realmente asociados a la esquizofrenia, con proporciones de probabilidad casi similares cercanas a 1,3 para esos diez o más polimorfismos y con un poder muy similar para identificar a cualquiera de ellos. Así, $10/100.000=10^{-4}$, y la probabilidad pre-estudio de que cualquier polimorfismo esté asociado con la esquizofrenia es también $R/(R + 1)=10^{-4}$. Supongamos también que el estudio tiene un poder del 60% para encontrar una asociación con una proporción de probabilidades de 1,3 a un nivel de $\alpha = 0,05$. Entonces se puede estimar que si se encuentra una asociación estadísticamente significativa con un valor p que apenas cruza el umbral de 0,05, la probabilidad post-estudio de que sea verdadera aumenta unas 12 veces en comparación con la probabilidad pre-estudio, pero es aún de solo 12×10^{-4}.

Ahora supongamos que los investigadores manipulan su diseño, su análisis y su informe para que más relaciones crucen el umbral de $p = 0,05$ aunque este no se habría cruzado con un diseño y un análisis que cumplieran perfectamente los estándares y con un perfecto informe integral de los resultados, siguiendo estrictamente el plan original del estudio. Esa manipulación se podría hacer, por ejemplo, excluyendo o incluyendo convenientemente ciertos pacientes o controles, con análisis *post hoc* de subgrupos, investigación de contrastes genéticos que no se especificaron originalmente, cambios en las definiciones de la enfermedad o de los controles, y diversas combinaciones de reporte selectivo o distorsionado de los resultados. Los paquetes comerciales de minería de datos se enorgullecen de su capacidad para obtener resultados estadísticamente significativos a través del dragado de datos. En presencia de un sesgo $u = 0,10$, la probabilidad post-estudio de que un hallazgo de investigación sea verdadero es de solo $4,4 \times 10^{-4}$. Además, incluso en ausencia de sesgo, cuando diez equipos de investigación independientes hacen experimentos similares en todo el mundo, si uno de ellos encuentra una asociación formalmente significativa en términos estadísticos, la probabilidad de que el resultado de investigación sea verdadero es de solo $1,5 \times 10^{-4}$, ¡apenas mayor que la probabilidad que tuvimos antes de emprender esta extensa investigación!

el valor predictivo de sus resultados de investigación. O las pruebas masivas orientadas al descubrimiento pueden dar como resultado tal cantidad de relaciones significativas que los investigadores tienen las suficientes para reportar e investigar más y así abstenerse del dragando y la manipulación de datos.

LA MAYORÍA DE LOS RESULTADOS DE INVESTIGACIÓN SON FALSOS EN LA MAYORÍA DE LOS DISEÑOS Y EN LA MAYORÍA DE LOS CAMPOS

En el marco descrito, un VPP superior al 50% es muy difícil de conseguir. El cuadro 4 presenta los resultados de las simulaciones que usan las fórmulas desarrolladas sobre la influencia del poder, la proporción relaciones verdaderas-no verdaderas y el sesgo, para varios tipos de situaciones que pueden ser características de diseños y campos de estudio específicos. Un hallazgo de una prueba aleatoria controlada, con un poder adecuadamente calibrado y bien realizada, que comienza con una posibilidad pre-estudio del 50% de que la intervención sea efectiva es eventualmente verdadero cerca del 85% de las veces. Se espera un desempeño muy similar en un meta-análisis confirmatorio de pruebas aleatorias de buena calidad: el sesgo potencial quizá aumente, pero el poder y las oportunidades antes de la prueba son más altos en comparación con una sola prueba aleatoria. A la inversa, un resultado meta-analítico de estudios no concluyentes, en los que el agrupamiento se usa para "corregir" el bajo poder de los estudios individuales, es probablemente falso si $R < 1{:}3$. Los resultados de investigación de ensayos clínicos de fase inicial con bajo poder serían verdaderos cerca de una a cuatro veces, o incluso menos frecuentes si el sesgo está presente. Los estudios epidemiológicos de naturaleza exploratoria se desempeñan aún peor, en especial cuando tienen poco poder, pero incluso los estudios epidemiológicos con buen poder pueden tener solo una probabilidad de uno a cinco de ser verdaderos, si $R = 1{:}10$. Por último, en la investigación orientada al descubrimiento con pruebas masivas, donde las relaciones probadas superan 1.000 veces a las verdaderas (p. ej., 30.000 genes probados, de los cuales 30 pueden ser los verdaderos culpables) (Ntzani y Ioannidis, 2003), el VPP de cada pretendida relación es sumamente bajo, aun con una considerable estandarización de los métodos estadísticos y de laboratorio, de los resultados y de los informes para minimizar el sesgo.

Cuadro 4
VPP de resultados de investigación para varias combinaciones de poder (1–β), proporción de relaciones verdaderas-no verdaderas (*R*) y sesgo (*u*)

1–β	*R*	*u*	Ejemplo práctico	VPP
0,80	1:1	0,10	PAC de poder adecuado con poco sesgo y 1:1 probabilidades pre-estudio	0,8500
0,95	2:1	0,30	Meta-análisis confirmatorio de PAC de buena calidad	0,8500
0,80	1:3	0,40	Meta-análisis de estudios pequeños no concluyentes	0,4100
0,20	1:5	0,20	PAC de poco poder, fase I/II bien realizada	0,2300
0,20	1:5	0,80	PAC de poco poder, fase I/II mal realizada	0,1700
0,80	1:10	0,30	Estudio epidemiológico exploratorio de poder adecuado	0,2000
0,20	1:10	0,30	Estudio epidemiológico exploratorio de poco poder	0,1200
0,20	1:1.000	0,80	Investigación exploratoria orientada al descubrimiento con pruebas masivas	0,0010
0,20	1:1.000	0,20	Como en el ejemplo anterior, pero con sesgo más limitado (más estandarizado)	0,0015

Los VPP estimados (valores predictivos positivos) se calculan suponiendo $\alpha = 0{,}05$ para un solo estudio.
PAC: prueba aleatoria controlada.
DOI: 10.1371/journal.pmed.0020124.t004

LOS RESULTADOS DE INVESTIGACIÓN REIVINDICADOS A MENUDO PUEDEN SER MEDIDAS PRECISAS DEL SESGO PREDOMINANTE

La mayor parte de la investigación biomédica moderna opera en áreas con muy baja probabilidad pre-estudio y post-estudio de hallazgos verdaderos. Supongamos que en un campo de investigación no hay hallazgos verdaderos por hacer. La historia de la ciencia nos enseña que, en el pasado, el empeño científico a menudo desperdició esfuerzos en campos que no producían ninguna información científica verdadera, al menos según nuestra comprensión actual. En ese "campo nulo", lo ideal sería esperar que todos los tamaños de efecto observados varíen al azar alrededor de 0 en ausencia de sesgo. El grado en que los resultados observados se desvían de lo que se espera solo por azar sería simplemente una medida pura del sesgo predominante.

Por ejemplo, supongamos que ningún nutriente o patrón dietético es realmente un determinante importante del riesgo de desarrollar un tumor específico. Supongamos también que la literatura científica ha examinado 60 nutrientes y afirma que todos ellos están relacionados con el riesgo de desarrollar este tumor con riesgos relativos de un rango de 1,2 a 1,4 para la comparación de los terciles superiores e inferiores de ingesta. Los tamaños de efecto pretendidos no miden entonces más que el sesgo neto involucrado en la elaboración de esta literatura científica. Los tamaños de efecto pretendidos son de hecho las estimaciones más precisas del sesgo neto. Se deduce incluso que entre "campos nulos", los campos que pretenden mayores efectos (a

menudo con las pretensiones acompañantes de importancia médica o de salud pública) son simplemente aquellos que han sostenido los peores sesgos.

En campos con VPP muy bajo, las pocas relaciones verdaderas no distorsionarían mucho esta imagen general. Incluso si algunas relaciones son verdaderas, la forma de la distribución de los efectos observados aún arrojaría una medida clara de los sesgos involucrados en el campo. Este concepto invierte por completo la manera de ver los resultados científicos. Tradicionalmente, los investigadores han visto con excitación efectos grandes y altamente significativos como signos de descubrimientos importantes. En realidad, los efectos demasiado muy grandes y muy significativos pueden ser signos de grandes sesgos en la mayoría de los campos de la investigación moderna. Deberían llevar a los investigadores a una cuidadosa reflexión crítica sobre lo que puede haber salido mal con sus datos, análisis y resultados.

Por supuesto, es probable que los investigadores que trabajan en cualquier campo se resistan a aceptar que todo el campo al que han dedicado su carrera sea un "campo nulo". Sin embargo, otras líneas de prueba, o avances en la tecnología y la experimentación, pueden llevar eventualmente al desmantelamiento de un campo científico. La obtención de medidas del sesgo neto en un campo también puede ser útil para obtener información de cuál podría ser el rango del sesgo en otros campos donde pueden existir métodos analíticos, tecnologías y conflictos similares.

¿CÓMO PODEMOS MEJORAR LA SITUACIÓN?

¿Es inevitable que la mayoría de los resultados de investigación sean falsos, o podemos mejorar la situación? Un problema importante es que resulta imposible saber con el 100% de certeza cuál es la verdad en cualquier pregunta de investigación. A este respecto, el patrón "oro" puro es inalcanzable. Sin embargo, hay varios enfoques para mejorar la probabilidad post-estudio.

Por ejemplo, evidencia con mayor poder; los grandes estudios o los meta-análisis con menor sesgo pueden ayudar, pues se acercan más al patrón "oro" desconocido. Aunque los grandes estudios pueden tener sesgos y estos se deberían reconocer y evitar. Además, es imposible obtener evidencia a gran escala para todos los millones y trillones de preguntas de investigación planteadas en la investigación actual. La evidencia a gran escala se debería centrar en preguntas de investigación

donde la probabilidad pre-estudio ya es muy alta, de modo que un hallazgo de investigación significativo lleve a una probabilidad post-comprobación que se considere bastante definitiva. La evidencia a gran escala también es particularmente indicada cuando pueda probar conceptos importantes en vez de preguntas estrechas y específicas. Un hallazgo negativo puede entonces refutar no solo una pretensión específica propuesta sino todo un campo o una parte considerable de ese campo. La selección del desempeño de estudios a gran escala con base en criterios estrechos, como la promoción comercial de un medicamento específico, es investigación desperdiciada. Además, se debe tener cuidado para que los estudios sumamente grandes tengan mayor probabilidad de encontrar una diferencia estadística formalmente significativa acerca de un efecto trivial que no sea significativamente diferente del efecto nulo (Lindley, 1957; Bartlett, 1957; Senn, 2001).

En segundo lugar, la mayoría de las preguntas de investigación son abordadas por muchos equipos, y es engañoso hacer énfasis en los hallazgos estadísticamente significativos de un equipo particular. Lo que importa es la totalidad de la evidencia. La disminución del sesgo mediante mejores estándares de investigación y la reducción de los prejuicios también pueden ayudar. Pero esto puede requerir un cambio de la mentalidad científica que puede ser difícil de lograr. En algunos diseños de investigación, los esfuerzos también pueden ser más exitosos con un registro anticipado de los estudios, por ejemplo, de los ensayos aleatorios (De Angelis, Drazen et al., 2004). El registro sería un reto para la investigación que genera hipótesis. Un tipo de registro o de interconexión de redes de bases de datos o de investigadores dentro de los campos puede ser más factible que registrar todos y cada uno de los experimentos que generan hipótesis. Pero aunque no veamos un gran avance en el registro de estudios en otros campos, podrían adoptarse más ampliamente los principios de desarrollo y adhesión a un protocolo de las pruebas aleatorias controladas.

Finalmente, en vez de perseguir la significancia estadística, deberíamos mejorar nuestra comprensión del rango de valores R –las probabilidades pre-estudio– donde operan los esfuerzos de investigación (Wacholder, Chanock, Garcia-C. et al., 2004). Antes de hacer un experimento, los investigadores deberían considerar la posibilidad de que crean estar probando una relación verdadera en vez de una no verdadera. A veces se pueden determinar unos altos valores conjeturados de R. Como se mostró más atrás, siempre que sean éticamente aceptables, se deben realizar grandes estudios con un sesgo mínimo sobre los hallazgos de investigación que se con-

sideran relativamente establecidos, para ver con qué frecuencia se confirman. Sospecho que varios "clásicos" establecidos no pasarán la prueba (Ioannidis, 2005).

No obstante, la mayoría de los nuevos descubrimientos seguirán siendo resultado de la investigación que genera hipótesis con bajas o muy bajas probabilidades pre-estudio. Deberíamos reconocer entonces que las pruebas de significancia estadística en el reporte de un solo estudio dan una imagen apenas parcial, sin saber cuántas pruebas se han realizado fuera del informe y en el campo relevante. A pesar de la abundante literatura estadística sobre múltiples correcciones de pruebas (Hsueh, Chen y Kodell, 2003), suele ser imposible descifrar cuánto han dragado los datos los autores informantes u otros equipos de investigación que han precedido a un hallazgo de investigación reportado. Incluso si fuese factible determinarlo, esto no nos informaría sobre las probabilidades pre-estudio. De modo que es inevitable hacer suposiciones aproximadas sobre cuántas relaciones se espera que sean verdaderas entre las que se prueban en los campos de investigación y diseño de investigación relevantes. El campo más amplio puede ofrecer una guía para estimar esta probabilidad en un proyecto de investigación aislado. También sería útil aprovechar las experiencias de sesgos detectados en otros campos vecinos. Aunque estas suposiciones serían muy subjetivas, serían muy útiles para interpretar las pretensiones de investigación y ponerlas en contexto.

REFERENCIAS BIBLIOGRÁFICAS

Altman, D. G. y Goodman, S. N. (1994). Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. *Journal of the American Medical Association*, *272*(2), 129-132.

Altman, D. G. y Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, *19*(4), 453-473.

Antman, E. M., Lau, J. et al. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *Journal of the American Medical Association*, *268*(2), 240-248.

Bartlett, M. S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, *44*(3-4), 533-534.

Chan, A. W., Hrobjartsson, A., Haahr, M. T. et al. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, *291*(20), 2457-2465.

Colhoun, H. M., McKeigue, P. M. y Smith, D. G. (2003). Problems of reporting genetic associations with complex outcomes. *Lancet*, *361*(9360), 865-872.

De Angelis, C., Drazen, J. M. et al. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine*, *351*(12), 1250-1251.

Golub, T. R., Slonim, D. K. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531-537.

Hsueh, H. M., Chen, J. J. y Kodell, R. L. (2003). Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. *Journal of Biopharmaceutical Statistics*, *13*(4), 675-689.

ICHE9 Expert Working Group (1999). Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine*, 18, 1905-1942.

Ioannidis, J. P. (2003). Genetic associations: False or true? *Trends in Molecular Medicine*, *9*(4), 135-138.

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*(2), 218-228.

Ioannidis, J. P. (2005). Microarrays and molecular research: Noise discovery? *Lancet*, *365*(9458), 454-455.

Ioannidis, J. P. y Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology*, *58*(6), 543-549.

Ioannidis, J. P., Evans, S. J. et al. (2004). Better reporting of harms in randomized trials: An extension of the Consort statement. *Annals of Internal Medicine*, *141*(10), 781-788.

Ioannidis, J. P., Haidich, A. B. y Lau, J. (2001). Any casualties in the clash of randomised and observational evidence? *British Medical Journal*, *322*(7291), 879-880.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. et al. (2001). Replication validity of genetic association studies. *Nature Genetics*, *29*(3), 306-309.

Kelsey, J. L., Whittemore, A. S. et al. (1996). *Methods in observational epidemiology*, 2.ª ed. Nueva York: Oxford University Press.

Krimsky, S., Rothenberg, L. S., Stott, P. y Kyle, G. (1998). Scientific journals and their authors' financial interests: A pilot study. *Psychother Psychosom*, *67*(4-5), 194-201.

Lawlor, D. A., Smith, D. G., et al. (2004). Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *Lancet*, *363*(9422), 1724-1727.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1-2), 187-192.

Marshall, M., Lockwood, A., Bradley, C. et al. (2000). Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. *British Journal of Psychiatry*, *176*(3), 249-252.

Michiels, S. Koscielny, S. y Hill, C. (2005). Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet, 365*(9458), 488-492.

Moher, D., Cook, D. J., Eastwood, S. et al. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet*, *354*(9193), 1896-1900.

Moher, D., Schulz, K. F. y Altman, D. G. (2001). The Consort statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*, *357*(9263), 1191-1194.

Ntzani, E. E. y Ioannidis, J. P. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment. *Lancet*, *362*(9394), 1439-1444.

Papanikolaou, G. N., Baltogianni, M. S., Contopoulos-I., D. G. et al. (2001). Reporting of conflicts of interest in guidelines of preventive and therapeutic interventions. *BMC Medical Research Methodology*, 1, 3.

Ransohoff, D. F. (2004). Rules of evidence for cancer molecular-marker discovery and validation. *Nature Reviews Cancer*, *4*(4), 309-314.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, *405*(6788), 847-856.

Senn, S. J. (2001). Two cheers for p-values. *Journal of Epidemiology and Biostatistics*, *6*(2), 193-204.

Sterne, J. A y Smith, D. G. (2001). Sifting the evidence-what's wrong with significance tests? *British Medical Journal*, *322*(7280), 226-231.

Stroup, D. F., Berlin, J. A., Morton, S. C. et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. Moose Group. *Journal of the American Medical Association*, *283*(15), 2008-2012.

Taubes, G. (1995). Epidemiology faces its limits. *Science*, *269*(5221), 164-169.

Topol, E. J. (2004). Failing the public health-rofecoxib, merck, and the FDA. *New England Journal of Medicine*, *351*(17), 1707-1709.

Vandenbroucke, J. P. (2004). When are observational studies as credible as randomised trials? *Lancet*, *363*(9422), 1728-1731.

Wacholder, S., Chanock, S., Garcia-C., M. et al. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, *96*(6), 434-442.

Yusuf, S., Collins, R. y Peto, R. (1984). Why do we need some large, simple randomized trials? *Statistics in Medicine*, *3*(4), 409-422.