Importancia de la potencia y la hipótesis en el valor p

Jordi Cortés[1,2], Martí Casals[3], Klaus Langohr[1], José Antonio González[1]

[1] Departamento de Estadística e Investigación Operativa, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

[2] Institut Universitari d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Barcelona, Spain

[3] CIBER de Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain
Servei d'Epidemiologia, Agència de Salut Pública de Barcelona, Barcelona, Spain

Dirección de correspondencia: jordi.cortes-martinez@upc.edu

Los lectores de Medicina Clínica conocen bien la importancia de definir bien el denominador de una proporción para estimar una probabilidad: "No es lo mismo la probabilidad de que un católico sea Papa que la de que un Papa sea católico" [1]. De forma similar, en un diagnóstico, no es lo mismo la probabilidad de que un enfermo dé positivo (*sensibilidad*), que la de que un caso que ha dado positivo esté enfermo (*valor predictivo de un positivo)*.

El valor p (o valor de p, o p-valor, o simplemente p) guarda cierta analogía con las probabilidades diagnósticas, ya que se define como la probabilidad de obtener un resultado tan significativo o más que el observado —dar positivo en la prueba diagnóstica— asumiendo cierta una hipótesis H: el paciente está sano. No obstante, a un investigador o a un clínico le puede resultar más interesante conocer el valor positivo de una prueba: cuán probable es una hipótesis H —que el paciente esté enfermo— habiendo observado unos resultados extremos.

En investigación, la replicabilidad y transparencia de un experimento es fundamental ya que nos acerca al rigor científico: cualquier investigador independiente debería poder reproducir y replicar nuestros resultados [2-4]. Veremos que esta replicabilidad está muy relacionada con la probabilidad previa de que la hipótesis alternativa sea cierta, con la potencia y la multiplicidad.

Una baja replicabilidad tiene importantes implicaciones. Por ejemplo, si la investigación previa en animales no se ha realizado con el máximo rigor, los estudios posteriores con voluntarios humanos serán inútiles [5]. Lo mismo sucede con la inversión material: *The Lancet* dedicó recientemente una serie de 5 artículos y 2 editoriales para denunciar el despilfarro de los recursos en investigación [6]. A todo ello contribuye el uso indiscriminado del valor p: Motulsky [7] enumera algunos de los

abusos: 1) reanalizar los datos de distintas maneras o aumentar el tamaño muestral hasta obtener un resultado significativo; 2) sobreinterpretar el valor p sin prestar atención al tamaño del efecto estimado; 3) realizar múltiples pruebas para la misma hipótesis sin corregir por multiplicidad, y 4) depender excesivamente de los errores estándar, frecuentemente malinterpretados. Todo ello conlleva que el valor p no sea necesariamente un indicador de la replicabilidad de los resultados [8].

Entonces, ¿cómo sabremos que un estudio es reproducible? La replicabilidad o consistencia será mayor cuanto mayor sea: 1) la base teórica y empírica para establecer la hipótesis; 2) la potencia del estudio, y 3) el control de la multiplicidad. Veámoslo con 2 aplicaciones.

Analogía

Empezaremos con una analogía a las pruebas diagnósticas: ¿qué influye en el valor predictivo de un resultado positivo? El lector puede encontrar en la dirección http://shiny-eio.upc.edu/bne/VPs una aplicación que proporciona los valores predictivos de los resultados de una prueba. Supongamos una anciana de 78 años con sospecha de embolismo pulmonar después de cirugía abdominal [**9**]. Los médicos coinciden en que la probabilidad de dicho trastorno en su intervalo de edad está en torno al 70%. Este valor representa la probabilidad *a priori*. Y aconsejan una gammagrafía pulmonar de ventilación-perfusión que tiene una sensibilidad del 40% y una especificidad del 98%. Con estos datos, haciendo uso de la aplicación (figura 1A), se deduce que la probabilidad *a posteriori* de la enfermedad es del 98%. Por tanto, la prueba, ha aportado información adicional incrementando la probabilidad del trastorno de un 70% previo ("a priori") a un 98% tras el resultado positivo ("a posteriori"). En otro ejemplo, un varón de 28 años con probabilidad a priori del 20%, tendría una probabilidad *a posteriori* del

83% (figura 1B). En ambos casos, la prueba aporta considerable evidencia. ¿Pero qué ocurriría si la especificidad fuese del 66% en vez del 98% (figuras 1C y 1D)? Entonces, las probabilidades *a posteriori* de la anciana y del joven habrían aumentado sólo del 70 al 73% y del 20 al 23%. Es decir, la prueba no habría añadido información.

¿Qué influye en la replicabilidad?

¿Qué ocurre con la credibilidad que merece un resultado científico? La aplicación está ahora en http://shiny-eio.upc.edu/bne/efectos. Supongamos que, en una determinada intervención, los estudios piloto —o de fase II, "proof of concept", "feasibility", según la terminología— han permitido avanzar desde la Investigación al Desarrollo. Imaginemos que estos estudios previos han permitido seleccionar intervenciones de tal forma que 3 de cada 4 son realmente eficaces (expectativa a priori de efecto real: 75%). Amparados en este escenario, se calcula el tamaño muestral necesario para garantizar, por ejemplo, una potencia del 80% para detectar tal efecto limitando el riesgo α (concluir que hay efecto cuando en realidad no lo hay) a un 5% unilateral ($1-\alpha=0.95$). En caso de que el experimento alcanzara la significación estadística, la aplicación obtiene que la confianza en un resultado positivo (proporción de efectos reales entre los resultados significativos) es del 98% (figura 2A), una cifra que permite esperar su futura replicación

Veamos ahora 3 maneras en las que podemos comprometer este buen resultado. En primer lugar, imaginemos que el equipo investigador no ha realizado un buen desarrollo previo de su hipótesis y pretenden ir directamente al estudio confirmatorio sobre el que pivotarán decisiones posteriores. Pongamos que estos falsos atajos bajan la expectativa del efecto a un 10%. En este caso, la confianza en un resultado positivo bajaría al 64% (figura 2B). En segundo lugar, si estos autores no disponen de muchos

casos, tendrán poca potencia. Una vez más, la aplicación muestra que para una potencia del 30%, la confianza en un resultado positivo bajaría al 40% (figura 2C). En este momento, ya es más probable que ese resultado significativo provenga en realidad de una intervención sin efecto. Finalmente, compruebe las consecuencias de atentar contra el tercer parámetro, α. Imaginemos que un investigador decide hacer muchas pruebas de hipótesis (lo que se conoce como "expedición de pesca") sobre diferentes variables, o bien sobre una misma variable repetida en diferentes ocasiones, ya sea a lo largo del tiempo o en diferentes condiciones de medida. Esta multiplicidad descontrolará el riesgo α y desvirtuaría la interpretación de los resultados. Si, por ejemplo, este investigador calcula 14 valores p en sendas pruebas independientes, la probabilidad de que simplemente por azar al menos 1 sea significativa es ligeramente mayor al 50%. En esta situación, la confianza en un resultado significativo bajaría hasta el 6.2% (figura 2D). Así, 1) sin un avance progresivo de la Investigación al Desarrollo (expectativa del efecto del 10%); 2) sin un diseño y tamaño muestral adecuados (30% de potencia), y 3) sin control del riesgo α (50%), sólo 6 de cada 100 intervenciones significativas tienen detrás un efecto real —las que podrían ser replicadas en el futuro.

Recomendaciones

Esta alarmante situación de 'desconfianza' en un resultado significativo fue descrita por John Ioannidis en 2005 [10], donde los ensayos clínicos y meta-análisis con una evidencia *a priori* de al menos del 50%, un α del 5% y una potencia del 80% conllevan una probabilidad *a posteriori* de evidencia real del 94%. Por otra parte, las expediciones exploratorias en busca de hipótesis al inicio de la I+D con una evidencia *a priori* que Ioannidis cuantifica por debajo del 1% y sin control de la probabilidad de error de tipo II (supongamos β=80%) pueden resultar en probabilidades del 3% de hallazgo real. En esta línea, la revista *Basic and Applied Social Psychology* ha decidido

eliminar los valores p de sus originales a partir de 2015 [11], originando gran controversia [12-14].

La metodología de un investigador en el proceso de I+D debe sostenerse en la transparencia absoluta sobre la situación de la hipótesis. No es ninguna vergüenza sugerir, en lugar de contrastar, una hipótesis. Al contrario, puede argumentarse que es más novedoso sugerir una nueva hipótesis que someter a contraste una ya conocida. A fin de cuentas, aunque se necesitaron varios viajes para abrir una vía comercial hacia las Américas, la auténtica novedad fue el primer viaje de Cristóbal Colón, en el que un hallazgo casual le permitió lanzar una hipótesis realmente innovadora.

En resumen, ¿jubilamos ya el valor p? Creemos que aún no es el momento, y que todavía es relevante. Sin embargo, sugerimos preparar el terreno, y de acuerdo con la transparencia [15], los autores deberían reportar 1) si la hipótesis era anterior a los resultados; 2) si tenían cálculo previo de potencia, y 3) si han respetado el riesgo α — por ejemplo, si han realizado una sola prueba y han sido fieles al plan previsto. Si algo no fuera así, deberían terminar diciendo que sus resultados sugieren pero no confirman su hipótesis, avanzando qué características deberían tener los estudios que permitirían contrastarla para así, poder progresar más rápidamente en la línea continua del I+D.

Agradecimientos: Agradecemos los comentarios recibidos por los miembros del grupo GRBIO que han permitido mejorar la calidad de este trabajo.

Financiación: No

Conflicto de intereses: No

Bibliografía

1. Senn S. Invalid inversion. Significance. 2013;10:40-2.
2. Leek JT, Peng RD. Reproducible research can still be wrong: Adopting a prevention approach. Proc Natl Acad Sci. 2015;112: 1645-6.
3. Von Elm E, Egger M. The scandal of poor epidemiological research. BMJ. 2004;329:868-9.
4. Altman DG. The scandal of poor medical research BMJ 1994;308:283.
5. Sooriakumaran P, Nyberg T, Akre O, Haendler L, Heus I, Olsson M, et al. Comparative effectiveness of radical prostatectomy and radiotherapy in prostate cancer: observational study of mortality outcomes. BMJ. 2014;348:g1502.
6. Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. The Lancet. 2014;383:166-75.
7. Motulsky HJ. Common misconceptions about data analysis and statistics. J Pharmacol Exp Ther. 2014;351:200-5.
8. Llobell JP, Pérez JFG, Navarro MDF. Significación estadística, importancia del efecto y replicabilidad de los datos. Psicothema. 2000;12(Suplemento):408-12.
9. Guyatt G, Rennie D. Guías para usuarios de literatura médica. Barcelona: Ars Medica; 2004.
10. Ioannidis JP. Why most published research findings are false. Plos Med. 2005; 2(8):e124.
11. Trafimow D, Marks M. Editorial. Basic Appl Soc Psych. 2015;37:1-2
12. Gelman A. Working through some issues. Significance. 2015;12: 33-5
13. Arenas C, Mestres F. Reliable software. BEIO. 2015;29:229-45

14. Schmidt AM. Banning null hypothesis significance testing. ISBA Bulletin. 2015;22;5
15. Catalá-López F, Hutton B, Moher D. Declaración de transparencia para las publicaciones científicas. Med Clin (Barc). 2014;142:554-5.

Figura 1: Un resultado positivo de una prueba diagnóstica puede aumentar la probabilidad de estar enfermo de forma notable (figuras A y B) o irrelevante (C y D).

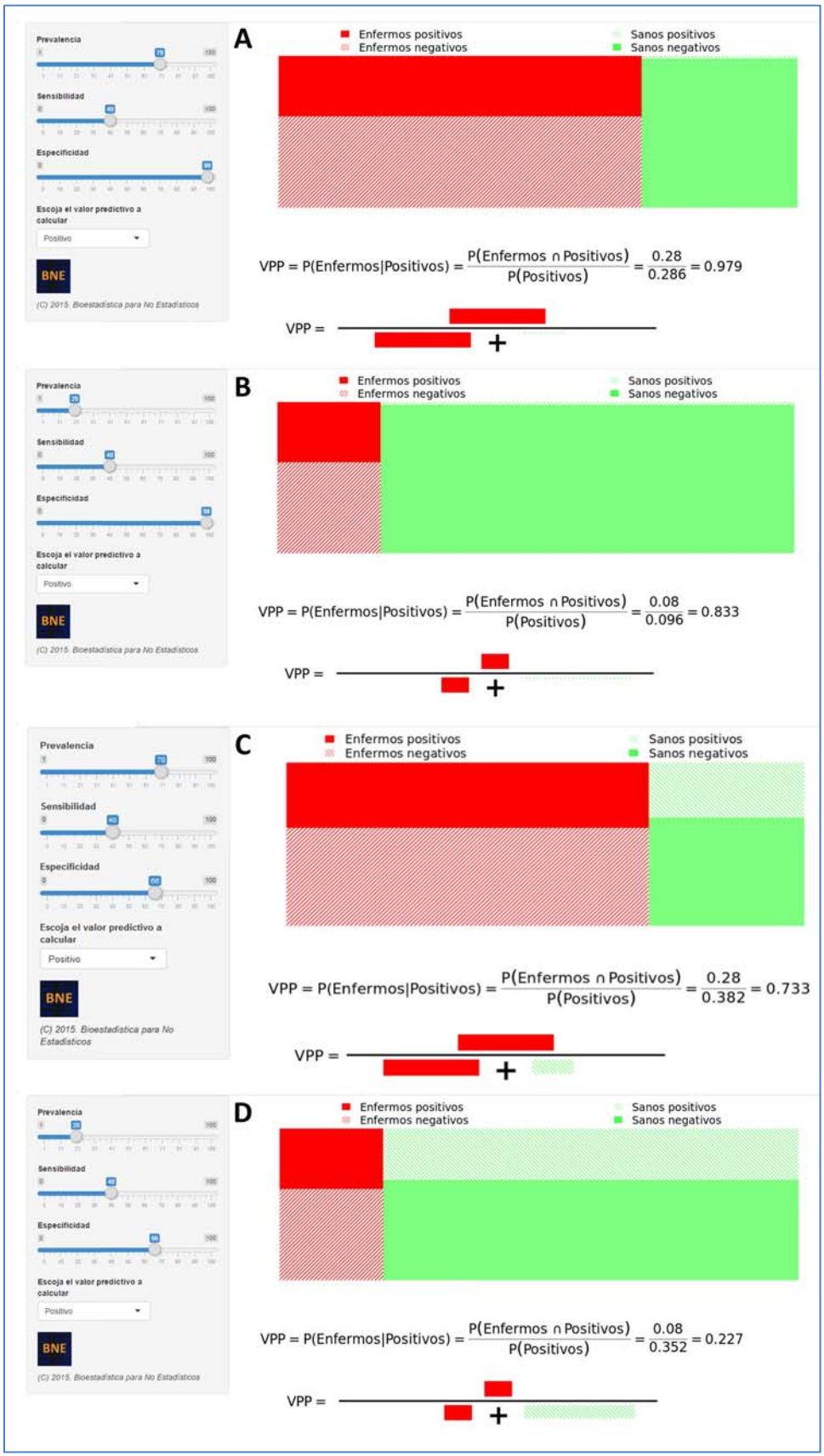

Figura 2: Probabilidad de replicación futura de un valor p significativo: en un estudio basado en una sólida investigación previa adecuadamente dimensionado (A); en las mismas condiciones pero sin sólida evidencia previa (B); además, sin adecuado dimensionamiento del tamaño que garantice su potencia (C) y con multiplicidad de análisis (D).