¿Significancia clínica o significancia estadística?

Clinical significance or statistical significance?

Agustín Ciapponi

Comentario Editorial

Resumen

El autor de este artículo explica como describir e interpretar el umbral clínico, el tamaño del efecto y la precisión de la estimación de los resultados de investigaciones biomédicas. Cuestiona la comunicación de la significancia estadística como única medida de reporte, que potencialmente podría conducir a errores de interpretación, y pone énfasis en la necesidad de informar al lector sobre la relevancia clínica o sanitaria de dichos hallazgos, explicando además la diferencia entre dos conceptos que pueden confundirse: 1) la ausencia de evidencia de efecto; 2) la evidencia a de ausencia de efecto.

Abstract

The author of this article explains how to describe and interpret the clinical threshold, the effect size and the estimation accuracy of the results of biomedical research. He disputes the reporting of statistical significance as the only measure, as potentially misleading, and emphasizes the need to inform the reader about the clinic or health relevance of these findings. The he explains the difference between two concepts that can be confused: 1) the absence of evidence effect, 2) the evidence of no effect.

Palabras clave: umbral clínico, significancia estadística, significancia clínica. **Key words:** clinical threshold, statistical significance, clinical significance.

Ciapponi A. Significancia estadística vs. relevancia clínica. Evid Act Pract Ambul Oct-Dic 2013; 16(4):122-125.

Descripción del tamaño del efecto y la precisión de la estimación

Es muy común que en el reporte de los resultados de las investigaciones biomédicas, especialmente las que comparan alternativas terapéuticas o preventivas, sean informadas las diferencias sólo a través de expresiones como "significancia estadística", "estadísticamente significativo" o sus negativos. Esto conlleva el peligro de que se confundan la significancia clínica y la estadística. Mucho más apropiado sería reportar las estimaciones centrales y sus intervalos de confianza, o cuando esto no sea posible, eventualmente los valores de "p", que si bien no son informativos del tamaño del efecto, al menos no están expuestos a confundir -no siempre de manera ingenua- conceptos estadísticos con los vinculados a la significancia clínica. Es decir que lo recomendable sería utilizar un lenguaje sencillo para describir el tamaño del efecto y la variabilidad o precisión de la estimación para que el lector interprete correctamente los resultados de una investigación.

Falta de evidencia de efecto y evidencia de ausencia de efecto

Un error de interpretación comúnmente cometido cuando la evidencia no es concluyente, es la confusión de dos conceptos cuyo significado es diferente:

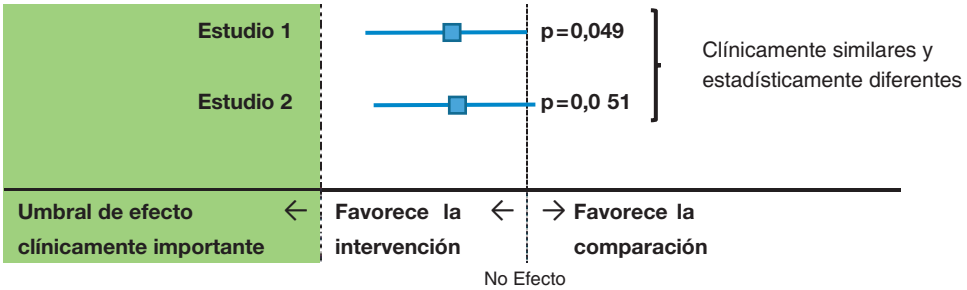
1) La "falta de evidencia de efecto", que implica que los resultados de las investigaciones que han abordado esta pregunta no han podido descartar el rol del azar en los resultados observados. Es decir que podría o no haber efecto pero no podemos estar seguros.

2) La "evidencia de ausencia de efecto", que se refiere a que tenemos un buen grado de certeza (aquí sí hay evidencia) de que realmente no hay efecto alguno[1], ya que la investigación en cuestión tuvo alta sensibilidad (o poder estadístico) para detectar posibles diferencias, por ejemplo 90%.

Por lo tanto, no siempre una evidencia no concluyente demuestra que una intervención carece de efecto. Tampoco debe confundirse la "significancia estadística" con el tamaño o la importancia de un efecto. Típicamente y por razones históricas se utiliza un punto de corte de 5% para establecer la significancia estadística. Esto significa que los resultados son considerados "estadísticamente significativos" si la probabilidad de que la diferencia observada sea consecutiva al azar es inferior al 5 % ($p < 0,05$). Sin embargo, un resultado con alta significancia estadística (p. ej.: $p < 0,000001$), implica que el hallazgo observado en la investigación en cuestión tiene una bajísima probabilidad de ocurrir por casualidad pero que podría ser clínicamente irrelevante. Por ejemplo, el caso de un analgésico que en promedio sólo mejora el dolor en un punto respecto al placebo (promedio de 74 con analgésicos y de 75 con placebo) en una escala que va de 0 (ausencia de dolor) a 100 (el máximo dolor imaginable). Por el contrario, cuando los resultados son "estadísticamente no significativos" tampoco se puede asumir automáticamente que no haya habido impacto. Suele utilizarse esta expresión cuando la probabilidad de que la diferencia observada ocurra por casualidad o azar es igual o superior al 5% ($p \geq 0,05$). Sin embargo, existen dos problemas con esta asunción. La primera es que el punto de corte de 5% es arbitrario y la segunda, que resultados estadísticamente no significativos, a veces mal etiquetados como "negativos", pueden o no ser concluyentes. La figura 1 muestra como el uso de los términos "estadísticamente no significativo" o "negativo" pueden ser engañosos.

Figura 1: problemas al clasificar resultados de estudios como "estadísticamente no significativos" o "negativos" (los cuadrados azules indican la estimación central del efecto y las líneas azules a cada lado del mismo, el intervalo de confianza del 95%).

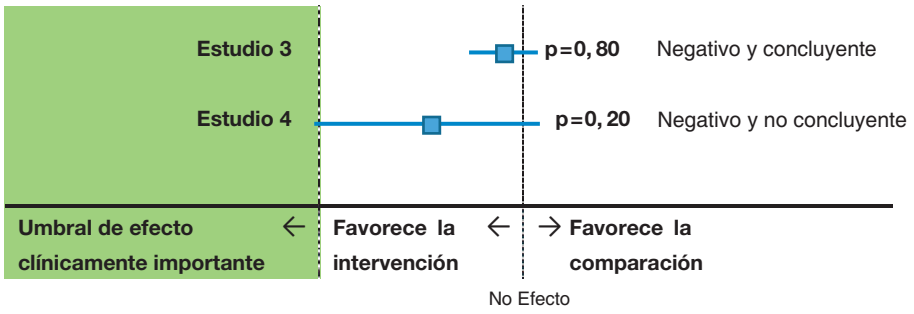
a. Clasificación basada en un punto de corte arbitrario.



Los resultados del estudio 1 son marginalmente diferentes a los del estudio 2, pero dado el punto de corte ($p < 0,05$), los resultados del estudio 1 se consideran "estadísticamente significativos" y los del estudio 2 "estadísticamente no significativos". Sin embargo, teniendo casi idéntico tamaño del efecto y similar impacto clínico (sin alcanzar el umbral de efecto clínicamente importante) no deberían considerarse diferentes entre sí.

Servicio de Medicina Familiar y Comunitaria del Hospital Italiano de Buenos Aires, Argentina y Centro Cochrane IECS-Instituto de Efectividad Clínica y Sanitaria. agustin.ciapponi@hiba.org.ar

b. Resultados "estadísticamente no significativos" pueden o no ser concluyentes en relación a un efecto clínicamente importante.



Los resultados del estudio 3 son "estadísticamente no significativos" o "negativos" pero además son concluyentes pues es extremadamente improbable que el azar haya impedido alcanzar el umbral de detección de un efecto clínicamente importante. Respecto del estudio 4, si bien sus resultados también son "estadísticamente no significativos" o "negativos", se trata de una situación en la que se puede concluir poco dado que podría haber algún efecto clínicamente importante que no haya podido ser demostrado.

Lamentablemente muchas veces los resultados son informados en forma tendenciosa, especialmente cuando se quiere resaltar la eficacia de alguna intervención. Por ejemplo, una tendencia "positiva" pero "estadísticamente no significativa"[§] (es decir, a favor de una intervención farmacológica), es interpretada como "prometedora", lo que puede resultar engañoso. Llamativamente, tendencias "negativas" de la misma magnitud, por el contrario, no suelen ser interpretadas como "señales de advertencia".

El umbral clínico

Ya explicada la cuestión del umbral de significancia estadística es necesario explicar brevemente el umbral clínico.

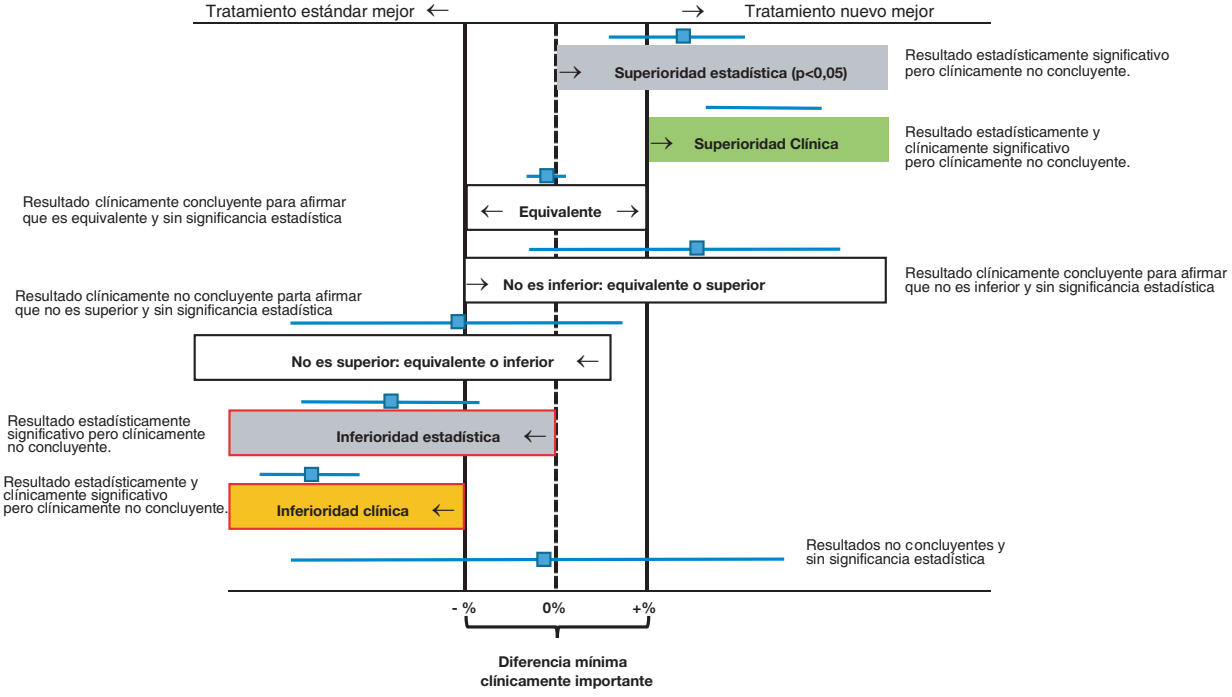
El primer paso es establecer la mínima diferencia clínicamente importante. Por ejemplo, si se asume que una diferencia relevante de descenso de presión arterial es un mínimo de 5 mmHg, cualquier diferencia menor a esa magnitud será considerada no importante. Vale destacar que esta diferencia no se basa en criterios estadísticos sino en conocimientos previos, en este caso, los que hayan demostrado que el impacto sobre la salud y/o sobre los costos en salud solo es apreciable cuando las diferencias son mayores a ese mínimo aceptado.

Por esta razón, en la sección métodos, los estudios deberían reportar el umbral que han establecido para determinar el tamaño muestral respecto del resultado primario**.

Un resultado será considerado clínicamente importante dependiendo de si su intervalo de confianza (IC95%) para a la diferencia entre ambos grupos cruza o no dicho umbral.

La figura 2 muestra una clasificación de diferentes tipos de resultados de estudios según umbrales estadísticos y clínicos, que como se ha mencionado, no necesariamente van de la mano.

Figura 2: clasificación de resultados de estudios según umbrales estadísticos y clínicos.



Los cuadrados azules indican la estimación central del efecto y las líneas azules a cada lado del mismo, el intervalo de confianza del 95%

§ Recordamos que esto se refiere a que existe una probabilidad más alta que la aceptada de que ese hallazgo haya sido por azar.

** Si no lo hubieran reportado, para poder valorar la importancia de la nueva información que aportan, ésta deberá ponerse en contexto a través de los antecedentes documentados en publicaciones anteriores.

Vale destacar que más difícil aún es intentar expresar los resultados de una investigación en un texto, ya que las posibilidades interpretativas se amplían considerablemente. En este contexto y para minimizar los malos entendidos, la Colaboración Cochrane ha adoptado un formato de reporte basado en una matriz que incluye información sobre el tamaño del efecto y la calidad de la evidencia (ver esta propuesta en tabla 1 y ejemplos de su aplicación en el cuadro 1)[2-4], basada en la clasificación GRADE[5-6].

Tabla 1: matriz utilizada por la Colaboración Cochrane, que integra el tamaño del efecto y la calidad de evidencia en la expresión de los resultados de una investigación clínica (para reportar un resultado reemplace dicha palabra por el específico de la investigación).

		Tamaño del efecto		
		Diferencia importante	**Diferencia pequeña**	**Diferencia mínima o nula[b]**
Calidad de evidencia[a]	**Alta**	Mejora, reduce, previene o conduce a un resultado	Mejora, reduce o conduce, ligeramente, a un menor o mayor resultado	Resulta en una diferencia mínima o nula (no influye en el resultado).
	Moderada	Probablemente mejora, reduce, previene o conduce a un resultado	Probablemente mejora o reduce o conduce, ligeramente, a un menor o mayor resultado	Probablemente conduce a una diferencia mínima o nula en el resultado
	Baja	Puede mejorar, reducir, prevenir o conducir a un resultado	Puede mejorar, reducir, prevenir o conducir , ligeramente, a un menor o mayor resultado	Puede conducir a una diferencia mínima o nula en el resultado
	Muy baja	Es incierto si la intervención mejora, reduce, previene o conduce a un cambio de resultado pues la calidad de la evidencia es muy baja		
Sin datos o estudios		Resultado no medido o reportado o no se encontraron estudios que hayan evaluado el impacto de la intervención en el resultado		

a Si bien el término calidad de la evidencia es el más conocido se está estudiando modificarlo por certeza de la evidencia. b En aquellos casos en los que "falta de evidencia de efecto" la expresión que haga referencia al resultado clínico debería agregar un comentario sobre el intervalo de confianza (p. ej. aplicado al estudio 4 de la figura 2: la clortalidona puede conducir a una diferencia mínima o nula en la mortalidad, pero no puede descartarse un descenso clínicamente importante).

Cuadro 1: un ejemplo de cómo expresar esta información

Quienes realizan largos vuelos podrían tener un mayor riesgo de desarrollo trombosis venosa profunda (TVP), especialmente los siguientes grupos de riesgo: ancianos, embarazadas, fumadores, personas con sobrepeso o con el antecedente de una TVP previa, trastornos de sangrado, enfermedad cardíaca o cáncer o cirugía reciente. Las medias compresivas ("medias de vuelo") aplican una suave presión, particularmente en el tobillo. Se usan durante todo el vuelo y pueden incrementar la circulación de la sangre en todo el cuerpo cuando su uso se combina con movimientos de las piernas. Vienen en diferentes tamaños y grados de compresión y deben ajustar apropiadamente.

En personas que realizan largos vuelos y comparado con no usar medias, el uso de medias compresivas reduce la incidencia de trombosis venosa profunda (TVP) asintomática, y probablemente conduce a una diferencia mínima o nula en la incidencia de trombosis venosas superficiales. Puede asociarse a una reducción de la hinchazón en las piernas. En los estudios analizados no hubo casos de TVP sintomática, embolismo pulmonar o muerte, lo que significa que estos estudios no pueden contestarnos si el uso de medias reduce la probabilidad de estos desenlaces. Solamente algunos estudios describieron los posibles problemas por usar medias. En éstos, los participantes no experimentaron ningún efecto secundario por utilizarlas. La tabla 2 resume estos hallazgos.

Tabla 2: estimación de lo que ocurre con y sin la utilización de medias elásticas

Resultados	Sin medias	Con medias	Comentarios	Calidad de la evidencia
Trombosis venosa profunda (TVP) con síntomas			Ningún participante desarrolló TVP sintomática en estos estudios	
Trombosis venosa profunda (TVP) sin síntomas	**Personas con bajo riesgo de TVP**			
	10 de 1000	1 de 1000 (Posiblemente tan bajo como 1 o tan alto como 3)		⊕⊕⊕⊕ Alta
	Personas con alto riesgo de TVP			
	30 de 1000	**3 de 1000** (Posiblemente tan bajo como 1 o tan alto como 8)	Medido después del vuelo, sobre una escala de 0, no edema, a 10, edema de máximo	⊕⊕⊕⊕ Alta
Trombosis de vena superficial	**13 de 1000**	**6 de 1000** (Posiblemente tan bajo como 2 o tan alto como 15)		⊕⊕⊕◯ Moderada
Hinchazón de pierna (edema)	El puntaje fue de **6 a 9**	El puntaje promedio fue **4,7 puntos menos** (Posiblemente 4,9 a 4,5 puntos menos)		⊕⊕◯◯ Baja

Los números dados son nuestra mejor estimación. Cuando fue posible, presentamos un rango porque hay una posibilidad del 95% de que el efecto verdadero se encuentre dentro de este rango.
Calidad alta: es muy improbable que investigaciones adicionales cambien nuestra confianza en la estimación del efecto. Calidad regular: es probable que investigaciones adicionales tengan un impacto importante sobre nuestra confianza en la estimación del efecto y puede cambiarla. Calidad baja: es muy probable que investigaciones adicionales tengan un impacto importante sobre nuestra confianza en la estimación del efecto y es probable que la cambie.

Este formato surge de estudios cualitativos y cuantitativos[7-9] y si bien ha implicado un gran avance hacia la correcta interpretación de los resultados de las investigaciones biomédicas, aún se requieren más estudios para seguir perfeccionándolos.

Recibido el 12/12/13 y aceptado el 10/01/14.

Referencias

1. Alderson P, Chalmers I. Survey of claims of no effect in abstracts of Cochrane reviews. BMJ. 2003-03-01 00:00:00 2003;326(7387):475.
2. Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. J Clin Epidemiol. Jun 2010;63(6):620-626.
3. Vandvik PO, Santesso N, Akl EA, et al. Formatting modifications in GRADE evidence profiles improved guideline panelists comprehension and accessibility to information. A randomized trial. J Clin Epidemiol. Jul 2012;65(7):748-755.
4. Santesso N, Glenton C, Rosenbaum S, et al. Evaluation of a Plain Language Summary template for Cochrane Reviews. Paper presented at: XV Cochrane Colloquium; 23-27/10/2007, 2007; São Paulo, Brasil.
5. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. Apr 2011;64(4):383-394.
6. Schünemann HJ, Oxman AD, Brozek J, et al. [Grading quality of evidence and strength of recommendations for diagnostic tests and strategies]. Chinese Journal of Evidence-Based Medicine. 2009(5):503-508. http://www.mrw.interscience.wiley.com/cochrane/clcmr/articles/CMR-13726/frame.html.
7. Glenton C, Santesso N, Rosenbaum S, et al. Presenting the results of Cochrane Systematic Reviews to a consumer audience: a qualitative study. Med Decis Making. Sep-Oct 2010;30(5):566-577.
8. Santesso N, Glenton C, Ciapponi A, Nilsen E, Pardo Pardo J, Rader T. Plain language summary format for Cochrane reviews: results of user testing. Paper presented at: XVI Cochrane Colloquium; 3-7/10/2008, 2008; Freiburg, Germany.
9. Santesso N, Glenton C, Ciapponi A, Stromme Nilsen E, Pardo Pardo J, Rader T. A new format for Plain Language Summaries: does it improve understanding, and is it useful and preferable? Paper presented at: XVII Cochrane Colloquium; 11-14/10/2009, 2009; Singapur.