Valor de *p* inferior a 0,05: ¿qué significa en realidad?

A pesar de que en un número cada vez mayor de estudios publicados se suscitan críticas sobre el uso de la simple significación estadística como medida de la repercusión clínica, sostenemos que este razonamiento no se ha incorporado de manera amplia en los estudios médicos pediátricos publicados. Esto es especialmente problemático, porque es necesaria una comprensión de las limitaciones del uso exclusivo de la significación estadística para evaluar los tratamientos con el objetivo de que los lectores de *Pediatrics* extraigan conclusiones precisas de los datos presentados en esta revista. En este comentario destacamos algunos de los aspectos relacionados con el problema complejo de evaluar los efectos del tratamiento y la importancia de usar la relevancia clínica, además del tradicional valor de *p*.

En la actualidad, el límite mágico de un valor de $p < 0,05$ sostiene una enorme importancia en lo que respecta a la aceptación de un manuscrito para su publicación o para la financiación de una solicitud de investigación o la aprobación de un nuevo fármaco por parte de la Federal Drug Administration. Sostenemos que si un tratamiento ha de ser útil para los niños, no es suficiente que los efectos de dicho tratamiento sean estadísticamente significativos; también han de ser de la magnitud suficiente para que sean significativos desde un punto de vista clínico. La evaluación de los resultados del tratamiento partiendo del valor de *p* exclusivamente es problemática por diversas razones. En primer lugar, con una muestra amplia es posible obtener un resultado estadísticamente significativo entre grupos a pesar de un efecto mínimo del tratamiento (es decir, pequeño tamaño del efecto). En segundo lugar, característicamente los pediatras malinterpretan unos resultados del estudio con valores de *p* más bajos como aquellos con efectos más potentes que los resultados con valores de *p* más altos. Es decir, la mayor parte de médicos creen que un resultado con un valor de $p = 0,002$ equivale a un efecto mucho mayor del tratamiento que un resultado con un valor de $p = 0,045$. Aunque esto es verdad si el tamaño de la muestra es igual en ambos estudios, no lo es si el tamaño de la muestra es mayor en el estudio donde se ha obtenido el menor valor de *p*. Esta confusión llega a ser particularmente preocupante cuando nos damos cuenta de que la mayor parte de estudios financiados por la industria farmacéutica se caracterizan por tamaños muy amplios de las muestras.

Para combatir la confianza excesiva en el valor de *p* cuando se examinan los efectos de los tratamientos, hacemos la recomendación siguiente: al examinar el artículo publicado sobre un ensayo clínico, el pediatra debe interesarse en responder a las tres preguntas básicas siguientes:

1. ¿Podrían los hallazgos del ensayo clínico ser exclusivamente el resultado de la casualidad? (es decir, significación estadística)
2. ¿Hasta qué punto es amplia la diferencia entre las variables primarias analizadas de los grupos del estudio? (es decir, consecuencias del tratamiento, tamaño del efecto)
3. ¿Es la diferencia de las variables primarias analizadas entre grupos significativa para el paciente? (es decir, relevancia clínica)

COMPRENSIÓN DE LA SIGNIFICACIÓN ESTADÍSTICA

Como resulta familiar para la mayor parte de lectores de *Pediatrics*, el valor de *p* es el método más utilizado de evaluación de la significación estadística de cualquier hallazgo. El origen del valor de *p* se remonta a 1925 cuando Sir Ronald A. Fisher sugirió por primera vez el uso de unos límites entre la significación y la falta de significación que se basaba en la probabilidad[1-3]. Arbitrariamente estableció este límite con una $p = 0,05$; donde *p* significa la probabilidad de que un hallazgo de interés se haya alcanzado por casualidad[1,2]. A pesar de que es bien conocido y se utiliza ampliamente el énfasis de Fisher en la prueba de la significación y el límite arbitrario de $p < 0,05$, es importante que los pediatras reconozcan que esta definición ha suscitado extensas críticas durante los 80 últimos años. Específicamente, se ha criticado esta estrategia porque no tiene en cuenta el tamaño y la relevancia clínica del efecto observado. Es decir, un efecto pequeño en un estudio con un gran tamaño de la muestra podría tener el mismo valor de *p* que un amplio efecto en un estudio con una muestra de pequeño tamaño.

En un intento de abordar algunas de las limitaciones del valor de *p*, algunos médicos han recomendado el uso de los intervalos de confianza[3] (IC). Sin embargo, es importante que el lector entienda que estas dos definiciones de significación estadística son esencialmente recíprocas[4]. Es decir, un valor de $p < 0,05$ es esencialmente lo mismo que un IC del 95% no superpuesto a 0. Sin embargo, los intervalos de confianza confieren ciertas ventajas en el sentido de que pueden usarse para estimar el tamaño de la diferencia entre grupos[5]. Por desgracia, esta estrategia no se utiliza de manera amplia en los estudios pediátricos publicados y, hoy día, los IC se usan sobre todo como criterio indirecto para probar la hipótesis más que para considerar la amplia variación del tamaño probable del efecto.

MÁS ALLÁ DEL VALOR DE *P*: TAMAÑOS DEL EFECTO

Proporcionando más información que los valores de *p* o que los intervalos de confianza, el grupo de estadísticos llamados "tamaños del efecto" son medidas de la magnitud de la diferencia entre grupos, estandarizadas mediante el control de la variación dentro de grupos. En otras palabras, mientras que el valor de *p* indica si es probable que la diferencia entre dos grupos de un estudio concreto ocurra exclusivamente por casualidad, el tamaño del efecto cuantifica la magnitud de la diferencia entre estos dos grupos. Puesto que el tamaño del efecto se basa en diferencias estandarizadas entre grupos y no en el tamaño de la muestra, evalúa mejor la potencia de la intervención. De particular pertinencia para los pediatras son los tamaños del efecto de tipo d, ya que son los utilizados principalmente para comparar dos grupos de tratamiento. El tamaño del efecto tipo d se define como la magnitud de la diferencia entre dos medias, dividida por la desviación estándar [(media del grupo de control – media del grupo de tratamiento)/desviación estándar del grupo de control]. Por lo tanto, el tamaño del efecto tipo d depende de la variación dentro del grupo de control y las diferencias entre el grupo de control y el de la intervención. Por convención, los tamaños del efecto tipo d que son de casi 0,20 se interpretan como pequeños, los de casi 0,50 se consideran "moderados" y los tamaños en los límites de 0,80 se consideran "amplios"[6]. Tamaños del efecto de otro tipo, el tipo de potencia del riesgo, incluyen los cocientes de probabilidad como la *odds ratio*, cociente de riesgo, diferencia de riesgo y reducción del riesgo relativo. Es probable que los médicos estén más familiarizados con esta estadística menos abstracta y puede ser útil comprender que la estadística de probabilidad es un tipo de tamaño del efecto. Existen una serie de tipos diferentes de tamaños del efecto pero una descripción de estos diversos tipos y fórmulas está fuera del alcance de este comentario, aunque el lector interesado puede consultar con diversos artículos de revisión donde se describen estos problemas[7,8].

MÁS TODAVÍA: RELEVANCIA CLÍNICA

Llegados a este punto, consideramos que es importante advertir a los lectores de *Pediatrics* que la magnitud del cambio (tamaño del efecto) no debe interpretarse como una indicación de significado clínico. En lugar de ello, la relevancia clínica de un tratamiento debe basarse en referencias externas proporcionadas por médicos y pacientes. Es decir, un pequeño tamaño del efecto puede seguir siendo clínicamente significativo y, del mismo modo, es posible que no lo sea un amplio tamaño del efecto. En realidad, se reconoce cada vez más que los métodos tradicionales utilizados, como las pruebas de significación estadística y los tamaños del efecto, deben complementarse con métodos para determinar los cambios clínicos significativos.

Aunque apenas se ha alcanzado un consenso sobre los criterios para estos estándares de eficacia, las definiciones más habituales de cambios clínicamente significativos incluyen: 1) los pacientes tratados obtienen una mejora estadísticamente fiable en las puntuaciones del cambio; 2) los pacientes tratados son indistinguibles desde un punto de vista empírico de una población sana después del tratamiento, o 3) cambios de como mínimo una desviación estándar (DE). El método más utilizado para evaluar la fiabilidad de las puntuaciones del cambio es el método de Jacobson-Truax en combinación con los puntos de corte clínicos[9]. Utilizando este método, se considera poco probable que el cambio sea producto de un error de determinación si el índice del cambio fiable (ICF) es de más de 1,96. Es decir, cuando el paciente obtiene una puntuación de cambio de más de 1,96 se puede suponer razonablemente que ha mejorado dicha puntuación.

La validez de cada uno de los métodos descritos previamente puede mejorarse todavía más estableciendo su validez externa (es decir, la perspectiva del paciente). Por ejemplo, Flor et al condujeron un metaanálisis a gran escala que evaluó la eficacia de un tratamiento multidisciplinario para el dolor crónico[10]. Los investigadores pusieron de relieve que, entre pacientes que recibieron la intervención, el dolor disminuyó en un 25% con un tamaño del efecto de 0,7. Aunque este hallazgo parece prometedor desde un punto de vista estadístico, el significado de los resultados cambia a la luz de los hallazgos de Colvin et al, que describieron que los pacientes sólo consideran un "tratamiento satisfactorio" cuando la mejora del dolor es del 50 %[11]. Por consiguiente, en este ejemplo, una reducción del 25% en las puntuaciones de dolor sería estadística pero no clínicamente significativa. Está claro que esta área en desarrollo merece una discusión adicional.

CONCLUSIONES

El problema de la significación clínica es de importancia primordial tanto para los investigadores como para los pediatras en ejercicio. Desde la vertiente de la investigación, es indispensable que los estudios evalúen sistemáticamente la significación tanto estadística como clínica para un progreso de nuestra comprensión de los efectos del tratamiento. Como tales, alentamos a los investigadores para que, como mínimo, documenten los tamaños del efecto, y, siempre que sea posible, incorporen validaciones externas del significado clínico. Desde la vertiente clínica, los pediatras deben entender la potencial desconexión entre la significación estadística y la clínica cuando toman decisiones sobre la adopción de nuevos tratamientos. La interpretación de cualquier hallazgo de investigación debe producirse en el contexto de la magnitud del cambio que ha tenido lugar y el significado clínico de los hallazgos.

AGRADECIMIENTOS

Este trabajo ha sido financiado en parte por los National Institutes of Health mediante el National Institute of Child Health and Human Development, subvención R01HD37007-02.

ZEEV N. KAIN, MD, MBA, Y JILL MACLAREN, PHD
Center for the Advancement of Perioperative Health y Department of Anesthesiology, Pediatrics, and Child Psychiatry, Yale University School of Medicine, New Haven, Connecticut, Estados Unidos.

BIBLIOGRAFÍA

1. Fisher RA. Statistical methods for research workers. 1.ª ed. Edimburgo, Escocia: Oliver and Boyd; 1925.
2. Fisher RA. Design of experiments. 1.a ed. Edimburgo, Escocia: Oliver and Boyd; 1935.

3. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med. 1986;105:429-35.
4. Feinstein AR. *P*-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol. 1998;51: 355-60.
5. Gardner MG, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. BMJ. 1986;292:746-50.
6. Cohen J. Statistical power analysis for the behavioral sciences. 2.ª ed. Mahwah, NJ: Lawrence Erlbaum; 1988.
7. Kirk R. Practical significance: a concept whose time has come. Educ Psychol Meas. 1996;56:746-59.
8. Snyder, P, Lawson S. Evaluating results using corrected and uncorrected effect size estimates. J Exp Educ. 1993; 61:334-49.
9. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991;59:12-9.
10. Flor H, Fydrich T, Turk DC. Efficacy of multidisciplinary pain treatment centers: a meta-analytic review. Pain. 1992; 49:221-30.
11. Colvin DF, Bettinger R, Knapp R, Pawlicki R, Zimmerman J. Characteristics of patients with chronic pain. South Med J. 1980;73:1020-3.