MEDICINA BASADA EN EVIDENCIA

Estudios randomizados interrumpidos precozmente por beneficio: ¿Muy buenos o muy malos?

Roberto Candia B[1,a], Luz María Letelier S[1,2,3], Gabriel Rada G[1,2].

Randomized trials stopped early for benefit: is it good or bad?

Al momento de analizar críticamente un estudio clínico randomizado (ECR), el lector debe decidir si el estudio es válido, es decir, si cumple con una serie de criterios que aseguren que la posibilidad de sesgo en el estudio es mínima[1]. En el último tiempo se han reconocido algunos elementos importantes, que no están considerados en los criterios clásicos de validez[2], pero que son lo suficientemente relevantes como para hacernos dudar de los resultados de un estudio. Esto es lo que ocurre cuando un estudio se detiene tempranamente por beneficio.

El objetivo del presente artículo es explicar el concepto de detención temprana por beneficio de un ECR, los fundamentos metodológicos y la evidencia de cómo ello puede inducir sesgo, haciendo que los resultados de estos estudios no sean del todo confiables.

CÁLCULO DEL TAMAÑO MUESTRAL

Cuando un investigador diseña un ECR para demostrar la utilidad de una nueva intervención, debe anticipar cuántos pacientes (n) requiere el estudio para demostrar diferencias entre el grupo que recibe la nueva terapia y el grupo control. Este proceso se denomina cálculo del tamaño muestral. Cuando el tamaño muestral (n de pacientes) en un ECR es muy pequeño, hay una alta probabilidad de que ocurra una de estas situaciones:

1. Que se encuentre una diferencia entre el grupo intervenido y el grupo de comparación (o control), sólo por efecto del azar (error alfa o tipo I).
2. Que no se encuentre una diferencia entre los grupos cuando ésta en realidad sí existe (error beta o tipo II).

La capacidad de un estudio de encontrar diferencias cuando éstas realmente existen (1-error beta), se denomina poder.

Para determinar el tamaño muestral adecuado, el investigador debe anticipar cuál es el beneficio o efecto que espera encontrar. Aunque las nuevas terapias obviamente no están totalmente demostradas, siempre existe «alguna idea» del beneficio o efecto clínico probable, ya sea por los resultados

[1]Unidad de Medicina Basada en Evidencia, Pontificia Universidad Católica de Chile.
[2]Departamento de Medicina Interna, Pontificia Universidad Católica de Chile.
[3]Servicio de Medicina, Hospital Sótero del Río. Santiago de Chile.
[a]Residente de Medicina Interna.

de otros estudios (randomizados u observacionales) o por fundamentos fisiopatológicos. Este «efecto clínico probable» es la diferencia estimada que va a existir entre el grupo control y el de intervención; es a partir de esa diferencia que se calcula el tamaño muestral[3,4].

Como se revisó en un artículo previo[5], mientras mayor es el n de la muestra, más estrecho es el intervalo de confianza (IC) en que se moverá el estimador puntual que resulta de un estudio. De esta forma, el n de la muestra necesario para demostrar diferencia estadísticamente significativa debe ser grande si se anticipan diferencias pequeñas entre los grupos, y al contrario, si las diferencias esperadas son grandes, el n necesario podría ser pequeño.

Ilustraremos lo anterior con un ejemplo: Supongamos que un investigador quiere demostrar una disminución de mortalidad al aplicar una nueva droga en pacientes con una enfermedad «X», y de acuerdo a los datos fisiopatológicos y estudios observacionales al respecto, estima que la diferencia será de alrededor de 5% (es decir, al aplicar la droga, la diferencia en la mortalidad entre los grupos intervención y control va a ser de 5%). El n necesario para demostrar diferencia debe ser suficiente para que el intervalo de confianza no toque la línea del no efecto, lo que en términos prácticos resulta en una diferencia estadísticamente significativa. Así, si estimamos una diferencia de 5%, un tamaño muestral pequeño generará un intervalo de confianza amplio que no es suficiente para demostrar una diferencia estadísticamente significativa (Figura 1).

Con un n mayor obtendremos un intervalo de confianza más angosto, por lo tanto, a pesar de estimar una diferencia pequeña (5%), ésta es significativa (Figura 2).

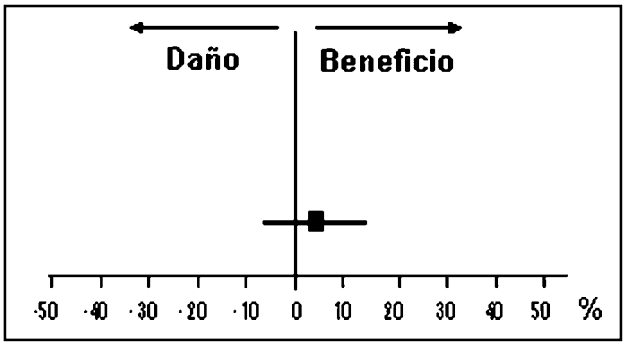

Figura 1. Diferencia estimada en 5%. Una muestra de n pequeño genera un IC amplio, que cruza la línea del no efecto (vertical) y, por tanto, no demuestra diferencias estadísticamente significativas. El n no es suficiente (error beta, poder insuficiente).

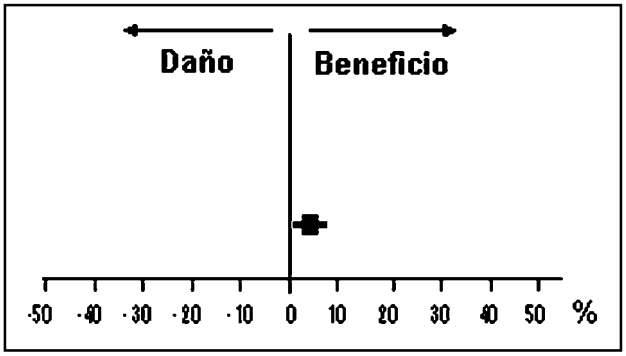

Figura 2. Diferencia estimada en 5%. Una muestra de gran tamaño genera un IC "estrecho", por tanto, logra demostrar diferencias significativas. El n es suficiente (poder adecuado).

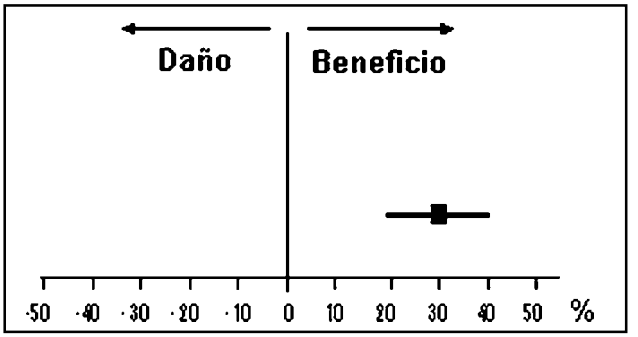

Figura 3. Diferencia estimada en 30%. Un n pequeño genera un IC amplio, pero de todas formas demuestra diferencias estadísticamente significativas. En esta situación un n pequeño puede ser suficiente.

En cambio, si las diferencias estimadas son grandes, por ejemplo, una diferencia de 30%, se necesita un n menor, ya que aunque el IC sea amplio, éste estará lo suficientemente «alejado» de la línea del no efecto como para demostrar una diferencia estadísticamente significativa (Figura 3).

En resumen, para calcular el tamaño muestral el investigador debe estimar el posible resultado y a partir de ese valor, calcular el n necesario para que el estudio tenga el poder suficiente (por convención se ha fijado en 80%) para que tales resultados sean estadísticamente significativos (habitualmente un valor p <0,05, también fijado por convención). A mayor n siempre existe menor posibilidad de error, ya que si realmente existen diferencias es más fácil encontrarlas (disminuye el error beta), y si se encuentran diferencias, con menor posibilidad éstas serán debidas al azar (disminuye el error alfa). Mientras mayor es el tamaño muestral de un estudio, con mayor probabilidad sus resultados se acercan al valor real.

¿Qué entendemos por detención precoz de un ECR y por qué ocurre?

El concepto de detención precoz se refiere a la interrupción del estudio antes de conseguir el tamaño muestral inicialmente estimado. No tiene ninguna relación con el tiempo de seguimiento[6].

Muchos investigadores, particularmente en estudios con tiempo de seguimiento muy largo, planifican analizar los resultados de sus estudios antes de completar el «n» inicialmente calculado, con la intención de observar tendencias hacia uno u otro grupo. Estos análisis se conocen como análisis interinos y les permiten decidir continuar o no con el estudio. Un estudio se detiene tempranamente por 2 razones:

1. El investigador pierde la esperanza de encontrar diferencias significativas.
2. El investigador encuentra diferencias de gran magnitud, que le parecen serían «improbables» por el azar (luego veremos que esto es un error). Estas diferencias pueden:
   - 2a. Mostrar daño con la intervención en estudio.
   - 2b. Mostrar beneficio con la intervención (detención temprana por beneficio).

Este último caso es el más relevante y el motivo de este artículo, ya que los estudios detenidos tempranamente por beneficio parecen exitosos, aparecen en revistas de alto impacto, son frecuentemente citados y se convierten en estudios fundamentales para guías de práctica clínica[7], a pesar de que sus resultados tienen un alto riesgo de error.

El riesgo de sobreestimar diferencias en ECR detenidos precozmente por beneficio

Cuando un ECR se detiene tempranamente, existe la posibilidad de pesquisar un beneficio «azarosamente alto». Supongamos que un ECR está evaluando un tratamiento con beneficio real, pero modesto. En las etapas iniciales de recolección de datos, sólo por azar, existe un alto riesgo tanto de

sobreestimar como de subestimar el efecto[7]. En algunas ocasiones esta sobreestimación va a ser de gran magnitud. A menor número de eventos, mayor es el rol que juega el azar en alejar el resultado del valor real.

Una forma de graficar este punto es con el ejemplo de la moneda: supongamos que queremos saber en forma experimental cuál es la probabilidad de que al lanzar una moneda al aire el resultado sea "sello". Para eso debemos lanzar la moneda en varias ocasiones y calcular esa probabilidad. En este caso, todos sabemos que la probabilidad de que el resultado sea sello es "50%" (ese es el valor real y el que nos interesa obtener). Si lanzo la moneda 10 veces, no es improbable que el resultado sea 8 veces "sello", aunque todos sepamos que ese 80% de probabilidad no es el valor real. Si detuviéramos el estudio en este punto podríamos concluir que la probabilidad que el resultado sea "sello" es de 80%, ya que ese es mi estimador puntual con un n de 10 lanzamientos. Si lanzamos la moneda 100 veces es menos probable que en 80 ocasiones el resultado sea "sello", y de hecho, a mayor número de repeticiones mayor es la probabilidad de que el valor observado sea cercano al real, es decir, 50%[8]. Si graficamos este ejemplo (Figura 4) vemos que cuando el experimento tiene un mayor n (en este caso, mayor número de lanzamientos de la moneda) más cercano está el estimador puntual (porcentaje de sellos) del valor real. A menor n, mayor es la variación del estimador puntual, sobreestimando o subestimando el efecto real de la intervención.

Este mismo efecto se observa al detener un estudio tempranamente, ya que sólo por el hecho de existir pocas repeticiones del evento (una muestra pequeña), con alta probabilidad el valor observado puede ser muy distinto del valor real, sólo por azar.

Si durante un análisis interino se demuestra un beneficio de gran magnitud, algunos investigadores decidirán detener el estudio aduciendo para ello razones éticas. Esto en principio parece razonable, dado que si no se detiene significaría no ofrecer una terapia "efectiva" a los pacientes en el grupo control. Como se ha explicado, los resultados de tales estudios pueden crear la falsa impresión de un efecto de gran magnitud, debido al azar y no por un efecto real. En ocasiones, cuando esto ha ocurrido, estudios posteriores que evalúan el mismo tratamiento, pero rechazan la detención temprana y completan muestras de mayor n, no demuestran el beneficio o este beneficio es de menor magnitud que en el estudio de bajo n, efecto que se conoce como "regresión a la verdad"[6].

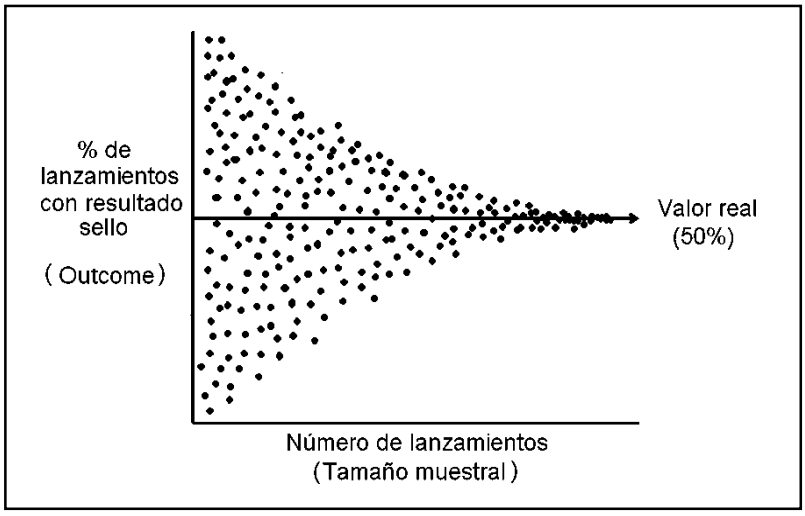

Figura 4. Figura que representa la variación del estimador puntual (resultado del estudio) en relación al tamaño muestral. Cada punto representa el resultado de un experimento.

Algunos ejemplos de estudios detenidos precozmente: resultados demasiado buenos para ser reales

Modelos de análisis estadísticos pueden rápidamente demostrar cómo los ECRs detenidos tempranamente sobreestiman efectos[9]; por otra parte, ECRs cuyos investigadores se niegan a detener sus estudios antes de lo planificado, también proveen de evidencia para demostrar este punto[7], como ocurre en el siguiente ejemplo:

Tifacogín (inhibidor del factor tisular) como tratamiento de pacientes con sepsis severa: el análisis interino al reclutar 722 pacientes mostró un beneficio de 10% en términos de mortalidad a los 28 días (demasiado bueno para cualquier tratamiento actual de sepsis), con valor p <0,006. Al completar el reclutamiento de pacientes (n =1.754) ese beneficio desapareció[10].

Evidencia similar se obtiene al repetir el diseño de ECR detenidos tempranamente, pero completando el tamaño muestral calculado[7] como ocurre con el uso de betabloqueo perioperatorio en que estudios iniciales mostraban beneficio de 90% en términos de reducción relativa de mortalidad al ser detenidos tempranamente[11], sin embargo, estudios posteriores (no detenidos tempranamente) y una revisión sistemática[12] demuestran que el beneficio, de ser real, sería de mucho menor magnitud.

Una revisión sistemática evaluó los ECRs detenidos tempranamente por beneficio[7]. De acuerdo a sus datos, de los estudios publicados entre 2000 y 2004, aproximadamente 1% fueron detenidos por esta causa y publicados en revistas de gran impacto. Se encontraron 143 estudios, de los cuales 55 fueron publicados en *New England Journal of Medicine*, 27 en *Lancet*, 6 en *JAMA*, 2 en *Annals of Internal Medicine*, 2 en *British Medical Journal* y 51 en otras revistas. La mayoría correspondían a estudios de cardiología, oncología y síndrome de inmunodeficiencia adquirida. En promedio estos estudios reclutaron 2/3 de los pacientes inicialmente planificados. Más de la mitad se detuvo con sólo 1 análisis interino. El riesgo relativo (RR) promedio para el *outcome* principal de cada estudio fue 0,53. Así, los estudios truncados precozmente por beneficio mostraban beneficios en términos de reducción de riesgo relativo de alrededor de 50%. Mientras menor era el número de eventos ocurridos, mayor era el beneficio estimado.

Considerando el conocimiento actual en biología y fisiopatología, asociado a la experiencia en relación a los beneficios de distintas terapias en humanos, la magnitud de estos beneficios es demasiado buena para ser real (existen pocas terapias que, en términos de riesgo relativo, disminuyan los eventos en 50%).

Consideraciones éticas y soluciones al problema

Los investigadores que después de un análisis interino encuentran diferencias estadísticas significativas a favor de la intervención se enfrentan a un dilema: ¿deben detener el estudio tempranamente para no privar de un tratamiento efectivo a los pacientes del grupo control? Como comentamos, el encontrar tales diferencias, muchas veces demasiado buenas, no reflejan en forma confiable el valor real. Detener un estudio en tales condiciones no es necesariamente ético, ya que a la larga muchos pacientes serán sometidos a un tratamiento cuyo beneficio está avalado por información que pudiese ser falsa. Algo distinto ocurre en los estudios detenidos por daño:

- Por una parte, sus datos habitualmente no son publicados (o al menos no tan difundidos), por lo tanto no son conocidos por la comunidad científica y los pacientes no son sometidos a terapias cuya utilidad está basada en datos no confiables.
- Por otro lado, si bien existe la posibilidad de cometer un error estadístico similar al que se comete al interrumpir por beneficio (atribuir un efecto que no existe a la intervención, sólo por el azar), la connotación ética en este caso es diferente. Si se detecta que muy probablemente los pacientes están siendo sometidos a un perjuicio dado por la intervención, tanto el compromiso adquirido con la persona que decidió voluntariamente participar en el estudio, como el principio clínico de "primero no hacer daño" (*primum non nocere*), obligan a detener el estudio.

Actualmente no existe una conducta totalmente aceptada para solucionar este problema. Algunos plantean simplemente no realizar análisis interinos, y completar todos los estudios, ya que esa es la información más confiable. Otros plantean una solución intermedia, estableciendo criterios de interrupción estrictos con tal de asegurar un número adecuado de eventos, por ejemplo:

* Establecer *a priori* comités externos independientes y "ciegos" para realizar los análisis interinos y decidir la detención del estudio. Esto debe estar explícito en la publicación del artículo, para que el lector pueda hacer su propia evaluación de la situación.
* Interrumpir los estudios sólo con diferencias altamente significativas (es decir, valores p muy pequeños). Esto para algunos es una mala solución, ya que si bien es menos probable que se detenga el estudio, sí se encuentra una diferencia, ésta de todas formas puede ser debida al azar, por ende la magnitud del efecto será necesariamente mucho mayor que el valor real.
* Asegurar un n mínimo de eventos: de acuerdo a la revisión sistemática antes mencionada[7], un número de eventos mayor a 200 se asoció a resultados más confiables.
* Establecer *a priori* un número limitado de análisis interinos, lo más alejados posible del inicio del estudio.

Dada la diversidad de posibles soluciones y la falta de acuerdo en la comunidad científica internacional sobre cómo enfrentar este problema, sugerimos al lector que cuando evalúe la validez de un estudio detenido tempranamente por beneficio, considere la adecuada entrega de información a este respecto: la descripción de los comités involucrados, cuál fue el criterio de detención, cuántos análisis interinos mediaron la interrupción, si éstos estaban planificados *a priori*. Así, el lector tendrá las herramientas suficientes para evaluar y decidir si aplica o no los resultados de tales estudios.

Conclusión

El tamaño muestral se relaciona con la posibilidad de error por azar en los resultados de un ECR. La interrupción temprana de estudios por beneficio habitualmente sobreestima el beneficio, sugiriendo que una intervención es efectiva cuando no lo es o que la magnitud del beneficio es mayor al real. Por ello, los resultados de estos estudios deben ser interpretados con cautela.

Referencias

1. Pantoja T, Letelier LM, Neumann I. El análisis crítico de la información publicada en la literatura médica. *Rev Méd Chile* 2004; 132: 513-5.
2. Peñalosa B, Candia R. ¿Por qué vale la pena randomizar un estudio de terapia? *Rev Méd Chile* 2004; 132: 1007-10.
3. Whitley E, Ball J. Statistics review 2: Samples and populations. *Critical Care* 2002; 6: 143-8.
4. Whitley E, Ball J. Statistics review 3: Hypothesis testing and P values. *Critical Care* 2002; 6: 222-5.
5. Candia R, Caiozzi G. Intervalos de confianza. *Rev Méd Chile* 2005; 133: 1111-15.
6. Pocock S, White I. Trials stopped early: too good to be true? *Lancet* 1999; 353: 943-4.
7. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M et al. Randomized trials stopped early for benefit: a systematic review. *JAMA* 2005; 294: 2203-9.
8. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians 2, Interpreting study results: confidence intervals. *CMAJ* 1995; 152: 169-73.
9. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials* 1989; 10 (4 Suppl): 209S-221S.
10. Abraham E, Reinhart K, Opal S, Demeyer I, Doig C, Rodriguez AL et al. Efficacy and safety of tifacogin (recombinant tissue factor pathway inhibitor) in severe sepsis: a randomized controlled trial. *JAMA* 2003; 290: 238-47.
11. Poldermans D, Boersma E, Bax JJ, Thomson IR, Van De Ven LL, Blankensteijn JD et al. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. *N Engl J Med* 1999; 341: 1789-94.
12. Devereaux PJ, Beattie WS, Choi PT, Badner NH, Guyatt GH et al. How strong is the evidence for the use of perioperative beta blockers in noncardiac surgery? Systematic review and meta-analysis of randomized controlled trials. *BMJ* 2005; 331: 313-21.