La ausencia de significación estadística en un ensayo clínico no significa equivalencia terapéutica

Localizador web
Artículo 43.617

Josep M. Argimon

Regió Sanitària Costa de Ponent. Servei Català de la Salut. L'Hospitalet de Llobregat. Barcelona.

«La ausencia de evidencia no es evidencia de ausencia.» Éste fue el título que Altman y Bland[1] eligieron para un artículo en el que alertaban sobre un error común en la bibliografía biomédica: interpretar un resultado estadísticamente no significativo, cuando se comparan dos tratamientos, como sinónimo de su equivalencia terapéutica.

La ausencia de significación estadística no implica que dos tratamientos tengan la misma eficacia o que puedan prescribirse indistintamente. En un estudio publicado recientemente se comparaba la eficacia de la warfarina y de la aspirina para prevenir las recurrencias de los accidentes vasculares cerebrales de origen isquémico[2]. Los investigadores diseñaron el estudio para poder demostrar una reducción relativa del riesgo del 30% en el grupo tratado con warfarina. Cuando finalizó el período de dos años de seguimiento, no se observaron diferencias estadísticamente significativas entre los dos grupos, y los autores llegaron a la conclusión de que la warfarina y la aspirina son opciones terapéuticas razonables en este tipo de situaciones clínicas. Sin embargo, la warfarina no sólo no consiguió reducir el riesgo en un 30%, sino que lo incrementó en un 13%, y el riesgo de hemorragias importantes tampoco fue inferior (2,22 por 100 pacientes/año en el grupo warfarina y 1,49 por 100 pacientes/año en el grupo aspirina). Además, tal como comentaban los autores en la discusión del artículo, la warfarina tiene un coste mayor y requiere una monitorización más estrecha. Con estos datos, ¿se puede considerar que la warfarina y la aspirina son alternativas terapéuticas razonables para prevenir los accidentes cerebrovasculares isquémicos recurrentes?

Generalmente, un valor de significación estadística mayor del 5% ($p > 0,05$) se etiqueta como no significativo. A los estudios con estos valores se les denomina negativos, un término que implica erróneamente que los tratamientos son iguales, cuando lo único que sucede es que no se ha podido demostrar una diferencia. Ambos conceptos son distintos. Un resultado que no alcance significación estadística quiere decir que, si en realidad no existe una diferencia, es probable que la observada en el estudio pueda haberse producido simplemente por variabilidad aleatoria. No debe interpretarse como indicativo de que no existe una diferencia en la realidad, sino tan sólo de que no puede descartarse esta posibilidad, en especial en los estudios que han incluido a pocos sujetos[3].

Contrariamente al objetivo de la mayoría de los ensayos clínicos, que es demostrar la mayor eficacia de un nuevo tratamiento comparada con la de el tratamiento estándar o un placebo, el objetivo de un ensayo clínico de equivalencia es demostrar que la eficacia del tratamiento experimental no es inferior a la del tratamiento estándar, es decir, que tienen efectos terapéuticos equivalentes. Este planteamiento tiene interés cuando el nuevo tratamiento ofrece ventajas adicionales como, por ejemplo, un menor número de efectos adversos, una administración más cómoda que facilite el cumplimiento de los pacientes, o un coste inferior al del tratamiento estándar[4,5].

Cuando un fármaco ha demostrado su eficacia ante una enfermedad grave y está bien establecido su uso para esta indicación, puede no ser ético realizar un ensayo clínico sobre un nuevo tratamiento experimental usando como control un placebo[6,7]. Además, una de las consecuencias de disponer de alternativas eficaces para el tratamiento de una enfermedad es la dificultad que tienen los nuevos tratamientos de demostrar una mayor eficacia que los ya existentes. Por estos motivos, cada vez se diseñan más estudios con el objetivo de demostrar que un nuevo tratamiento es tan bueno, o casi, como el prescrito habitualmente y, en teoría, presenta alguna ventaja adicional.

Se han dedicado muchos artículos y libros a la explicación del diseño y análisis de los ensayos clínicos que tratan de demostrar la superioridad de un tratamiento sobre otro, pero el planteamiento y análisis de los estudios de equivalencia ha merecido menos atención. En este artículo se abordan dos temas relacionados con el diseño de este tipo de estudios, fundamentales para poder interpretar correctamente sus resultados: el cálculo del tamaño de la muestra necesaria y la estrategia de análisis de los datos.

Cálculo del tamaño de la muestra

Aunque existen varios abordajes para demostrar que dos tratamientos tienen efectos terapéuticos equivalentes, uno de los más útiles e intuitivos para el clínico consiste en definir, cuando se diseña el estudio, un intervalo de valores de diferencias entre tratamientos (desde –d hasta +d), todos ellos compatibles con una diferencia sin importancia clínica. Si los límites del intervalo de confianza de la diferencia, calculado una vez que ha finalizado el estudio, se encuentran dentro del intervalo de valores de la diferencia predefinido, se puede concluir que ambos tratamientos tienen efectos terapéuticos equivalentes[8,9].

A diferencia de los estudios de superioridad, en los que para calcular el tamaño de la muestra se decide cuál es la mínima diferencia clínicamente relevante que se desea detectar, definida ésta como la diferencia entre los resultados que induciría a adoptar la mejor de las dos terapias[10], en los estudios de equivalencia se define la máxima diferencia (d) que se puede aceptar entre ambos tratamientos para considerar que son equivalentes. El procedimiento que se sigue es el de utilizar este valor como la diferencia máxima que se desea detectar en el cálculo del número de sujetos. De esta forma, si existe una diferencia real igual o mayor, el estudio tiene las suficientes garantías (potencia estadística) para detectarla y descartar la equivalencia terapéutica entre los tratamientos.

Correspondencia: Sr. J.M. Argimon.
Avda. de la Granvia, 8-10, 5.ª planta.
08902 L'Hospitalet de Llobregat. Barcelona.
Correo electrónico: jargimon@rscp.scs.es

Recibido el 24-1-2002; aceptado para su publicación el 13-2-2002.

TABLA 1

Fórmula para el cálculo del tamaño de la muestra en un estudio de equivalencia (variable cualitativa) y proporciones grandes (en general > 0,05)

$$n = \frac{2\,(Z_{\alpha/2} + Z_{\beta})^2\,P\,(1-P)}{d^2}$$

n: número de sujetos necesarios en cada uno de los grupos; $Z_{\alpha/2}$: valor de Z correspondiente al riesgo α fijado en un contraste bilateral (si es de 0,05 el valor de Z correspondiente es de 1,96); Z_{β}: valor de Z correspondiente al riesgo β fijado (para un riesgo de 0,10 el valor de Z correspondiente es 1,282); P: proporción correspondiente al grupo control; d: diferencia máxima tolerable entre la eficacia de ambos tratamientos para asumir que son equivalentes.

TABLA 2

Número de sujetos necesarios en cada grupo en estudios de equivalencia cuando la variable de respuesta es dicotómica

p	d				
	0,01	0,02	0,03	0,04	0,05
0,05	9.985	2.496	1.109	624	399
0,10	18.919	4.730	2.102	1.182	757
0,15	26.802	6.700	2.978	1.675	1.072
0,20	33.634	8.408	3.737	2.102	1.345
0,25	39.415	9.854	4.379	2.463	1.577
0,30	44.144	11.036	4.905	2.759	1.766
0,35	47.823	11.956	5.314	2.989	1.913
0,40	50.451	12.613	5.606	3.153	2.018
0,45	52.027	13.007	5.781	3.252	2.081
0,50	52.553	13.138	5.839	3.285	2.102
0,55	52.027	13.007	5.781	3.252	2.081
0,60	50.451	12.613	5.606	3.153	2.018
0,65	47.823	11.956	5.314	2.989	1.913
0,70	44.144	11.036	4.905	2.759	1.766
0,75	39.415	9.854	4.379	2.463	1.577
0,80	33.634	8.408	3.737	2.102	1.345
0,85	26.802	6.700	2.978	1.675	1.072
0,90	18.919	4.730	2.102	1.182	757
0,95	9.985	2.496	1.109	624	399

p: proporción esperada de éxito o curación en el grupo control; d: diferencia máxima tolerable entre la eficacia de ambos tratamientos para asumir que son equivalentes. Se asume un riesgo α de 0,05 (contraste bilateral) y un riesgo β de 0,10 (potencia estadística de 0,90).

Supongamos que se lleva a cabo un estudio hipotético para evaluar la eficacia de una pauta monodosis para el tratamiento de las infecciones urinarias frente a la pauta habitual de 7-10 días, cuya eficacia se sitúa en alrededor del 90%. La monodosis tiene un coste menor y facilita el cumplimiento del paciente, por lo que interesa determinar si puede considerarse equivalente terapéuticamente a la pauta habitual de 7-10 días. Los investigadores preestablecen que ambos tratamientos se considerarán equivalentes si la diferencia de eficacia entre ellos no supera el 5%. Si se aceptan los niveles de error α del 5% (hipótesis bilateral) y ß del 10% (potencia estadística del 90%), se estima, después de aplicar la fórmula aproximada de la tabla 1, que son necesarios 757 sujetos por grupo. La misma información se obtendrá, y sin necesidad de cálculos complejos, usando la tabla 2. Existen también fórmulas[8] y tablas[3] cuando la variable de respuesta es continua.

La elección y aplicación del intervalo de valores de la diferencia entre tratamientos que van a considerarse para definir la equivalencia siempre es difícil y controvertida. El concepto de equivalencia requiere descartar pequeñas diferencias en la respuesta a las intervenciones. El problema reside en definir qué se entiende por «pequeñas». Si se desea determinar con un alto grado de confianza la existencia de una equivalencia fijando diferencias muy pequeñas, se requerirá un número muy elevado de sujetos[3]. Un ejemplo se encuentra en dos ensayos clínicos sobre la eficacia del tratamiento trombolítico en el infarto agudo de miocardio. En el estudio INJECT[11] (International Joint Efficacy Comparison of Thrombolitics) los investigadores prefijaron como límite máximo una diferencia absoluta de mortalidad del 1%, una decisión basada en el hecho de que ésta fue la diferencia absoluta de mortalidad observada en el estudio GUSTO[12] (Global Utilization of Streptokinasa and Tissue Plasminogen Activator for Occluded Coronary Arteries). En el estudio COBALT[13] (Continuous Infusion versus Double-Bolus Administration Alteplase) los investigadores decidieron que la diferencia máxima de mortalidad que podía aceptarse para concluir que los tratamientos eran equivalentes era del 0,40%, dado que éste era el valor del límite inferior del intervalo de confianza alrededor del 1% de diferencia absoluta de mortalidad observado en el estudio GUSTO[12].

En el estudio INJECT[11] se observó una mortalidad del 9,53% en un grupo y del 9,02% en el otro (diferencia absoluta del 0,51%, menor del 1% prefijada por los investigadores y estadísticamente no significativa) y se concluyó que ambos tratamientos eran equivalentes. En el estudio COBALT[13] se asignaron 7.169 pacientes, y la diferencia en las tasas de mortalidad a los 30 días entre los grupos fue del 0,44% (7,98 frente al 7,53%), superior a la diferencia prefijada por los investigadores, por lo que la conclusión fue que ambas intervenciones no podían considerarse equivalentes.

La máxima diferencia que puede aceptarse entre los distintos tratamientos debe ser fijada en términos realistas, ya que tiene gran influencia sobre el tamaño de la muestra y, en consecuencia, sobre la viabilidad del estudio. En un editorial[14] que acompañaba a la publicación del estudio COBALT[13], se estimó que si realmente se considera que un 0,40% en la mortalidad a los 30 días es la máxima diferencia que puede aceptarse clínicamente para demostrar una equivalencia terapéutica, con una potencia estadística del 0,80%, se necesitaría incluir a más de 50.000 pacientes por grupo.

Es imprescindible que el intervalo de diferencias entre los tratamientos esté definido explícitamente antes de iniciar el estudio, incluso cuando la diferencia observada es muy pequeña, ya que de lo contrario la decisión de equivalencia deja de tener un componente de razonamiento clínico para convertirse simplemente en un tema de significación estadística[15]. Ésta no es una práctica habitual. En una revisión reciente se estimó que sólo el 23% de los estudios que afirmaban la equivalencia terapéutica entre dos intervenciones había prefijado los límites en la diferencia de eficacia que serían aceptables para decidir la equivalencia[16].

Estrategia de análisis

En un ensayo clínico suelen presentarse ciertas situaciones que obligan a considerar si determinados sujetos u observaciones deben ser excluidos del análisis. Según la actitud que se adopte ante estas situaciones, las conclusiones del estudio pueden ser diferentes.

La principal característica de los ensayos clínicos es la asignación aleatoria de los sujetos, que tiende a asegurar la comparabilidad entre los grupos de estudio. Dado que las pérdidas de seguimiento y retiradas en un estudio difícilmente se producen al azar, cualquier exclusión de sujetos del análisis puede alterar dicha comparabilidad, generar sesgos y conducir a conclusiones erróneas. Si, por ejemplo, algunos sujetos de un grupo no finalizan el estudio porque presentan acontecimientos adversos a un tratamiento, su exclusión del análisis conducirá probablemente a un sesgo favorable a este tratamiento. También puede conducir a sesgos la exclusión de sujetos que no finalizan el estudio

porque se han «curado», o bien porque han requerido tratamiento adicional por ineficacia de la intervención recibida[17]. Además, la exclusión de sujetos u observaciones del análisis disminuye la potencia estadística del estudio, es decir, se reduce la capacidad para detectar una diferencia o asociación de interés, ya que el número de individuos u observaciones que se tienen en cuenta es inferior al inicialmente previsto, lo que dificulta la obtención de resultados estadísticamente significativos. Por ello es fundamental que cualquier ensayo clínico se diseñe rigurosamente, se ejecute según lo acordado en el protocolo y se efectúe un seguimiento estricto de los sujetos, para obtener mediciones válidas en el máximo número de individuos y para minimizar las desviaciones del protocolo que pueden sesgar los resultados del estudio.

Con el fin de preservar la comparabilidad entre los grupos, la opción más válida para analizar un ensayo clínico cuyo objetivo es probar la superioridad de una intervención sobre otra consiste en evaluar a todos los pacientes incluidos en el estudio según el principio denominado análisis según intención de tratar *(intention-to-treat analysis)* o según asignación aleatoria (as-*randomized*), según el cual se analiza a todos los pacientes como pertenecientes al grupo al que fueron inicialmente asignados, independientemente de cualquier desviación del protocolo que se haya producido. Ésta es la opción más conservadora, ya que dificulta la obtención de un resultado estadísticamente significativo y, además, refleja lo que ocurre realmente en la práctica clínica[17].

Sin embargo, en los ensayos clínicos de equivalencia, esta modalidad de análisis deja de ser conservadora, precisamente porque permite concluir con más facilidad que los resultados no son estadísticamente significativos y, por tanto, inferir que las alternativas en estudio son terapéuticamente equivalentes[8,9]. Una alternativa al análisis según intención de tratar es la de comparar solamente los pacientes que han sido asignados a un grupo, que cumplen con todos los criterios de inclusión y exclusión, que han recibido el tratamiento correspondiente y que han sido seguidos hasta el final del estudio. Esta estrategia se conoce como análisis de casos válidos o según protocolo *(per-protocol)* y se supone que aplicándola, generalmente (aunque no siempre), aumentarán las diferencias entre los tratamientos, lo que dificultará poder concluir que dos tratamientos son equivalentes[8,9], como se demuestra en una revisión de 11 ensayos clínicos diseñados para demostrar la equivalencia terapéutica entre los esteroides y los betaagonistas, ambos inhalados, para el tratamiento del asma, donde los intervalos de confianza siempre fueron más amplios cuando los datos se analizaron según protocolo siguiendo la modalidad de análisis según intención de tratar[18].

En los estudios de equivalencia está justificado efectuar las dos modalidades de análisis. En la mayoría de las ocasiones los resultados de ambas estrategias permitirán llegar a la misma conclusión acerca de si los tratamientos son o no equivalentes. Si no sucede así, hay que investigar los motivos de las diferencias y analizar cuidadosamente los subgrupos de pacientes que se han desviado del protocolo. De todos modos, si dos tratamientos tienen patrones diferentes de pérdidas, retiradas u otras desviaciones del protocolo, significa que no son enteramente equivalentes.

En resumen, el lector de un artículo cuya conclusión es que dos tratamientos son equivalentes debe fijarse, en primer lugar, en si el objetivo inicial era probar la equivalencia o bien si esta conclusión se basa en que el resultado del análisis de un estudio diseñado para probar la superioridad de un tratamiento no ha sido estadísticamente significativo. En segundo lugar, debe fijarse en si la diferencia máxima prefijada por los investigadores en el protocolo de su estudio es razonable desde el punto de vista clínico y, por último, si la estrategia de análisis ha sido adecuada y el número de individuos que se desvían del protocolo es escaso. Llegados a este punto, si se cumplen estos requisitos, se procederá a evaluar si otros factores son lo suficientemente persuasivos como para recomendar el uso del nuevo tratamiento.

Agradecimientos

A la Sra. Victòria Ferrando y al Sr. Joan Vila por sus comentarios y aportaciones.

REFERENCIAS BIBLIOGRÁFICAS

1. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.
2. Mohr JP, Thompson JLP, Lazar RM, Levin B, Sacco RL, Furie KL, et al. A comparison of warfarin and aspirin for the prevention of recurrent ischemic stroke. N Engl J Med 2001;345:1444-51.
3. Argimon JM, Jiménez Villa J. Métodos de investigación clínica y epidemiológica. 2.ª ed. Madrid: Harcourt, 2000; p. 114.
4. Bristol DR. Clinical equivalence. Biopharm Stat 1999;9:549-61.
5. Hatala R, Holbrook A, Goldsmith CH. Therapeutic equivalence: all studies are not created equal. Can J Clin Pharmacol 1999;6:9-11.
6. Temple RT, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: Ethical and scientific issues. Ann Intern Med 2000;133:455-63.
7. Ellenberg SS, Temple RT. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: Practical issues and specific cases. Ann Intern Med 2000;133:464-70.
8. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996;313:36-9.
9. Massel D. Similar, the same or just not different: a guide for deciding whether treatments are clinically equivalent. Can J Cardiol 1999;15: 556-62.
10. Marrugat J, Vila J, Pavesi M, Sanz F. Estimación del tamaño de la muestra en la investigación clínica y epidemiológica. Med Clin (Barc) 1998; 111:267-76.
11. International Joint Efficacy Comparison of Thrombolytics. Randomised, double-blind comparison of reteplase double-bolus administration with streptokinase in acute myocardial infarction: trial to investigate equivalence. Lancet 1995;346:329-36.
12. The GUSTO investigators. An international randomized trial comparing four thrombolytic strategies for acute muyocardial infarction. N Engl J Med 1993;329:673-82.
13. The continuous Infusion versus Double-Bolus Administration of Alteplase (COBALT) Investigators. A comparison of continuous infusion of alteplase with double bolus administration for acute myocardial infarction. N Engl J Med 1997;337:1124-30.
14. Ware JH, Antman EM. Equivalence trials. N Engl J Med 1997;337:1159-61.
15. Argimon JM. El intervalo de confianza: algo más que un valor de significación estadística [en prensa]. Med Clin (Barc) 2002.
16. Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence? Ann Intern Med 2000;132:715-22.
17. Bakke OM, Carné X, García Alonso F. Ensayos clínicos con medicamentos. Fundamentos básicos, metodología y práctica. Barcelona: Doyma, 1994; p. 201-14.
18. Ebbutt AF, Frith L. Practical issues in equivalence trials. Stat Med 1998;17:1691-701.