Metodología para los estudios clínicos de superioridad frente a los de equivalencia y estudios de no inferioridad. Una revisión aplicada

Martin Rosas-Peralta,[a,g] Luis Efrén Santos-Martínez,[b]
José-Antonio Magaña-Serrano,[c] Jesús Salvador Valencia-Sánchez,[d]
Martin Garrido-Garduño,[e] Gilberto Pérez-Rodríguez[f]

Methodology for superiority versus equivalence and non-inferior clinical studies. A practical review

Physicians should always remember that a negative result in a superiority trial never would prove that the therapies under research are equivalent; more often, there may be a risk of type 2 (false negative) error. Equivalence and not inferiority studies demand high standards to provide reliable results. Physicians should take into account above all that the equivalence margins tend to be too large to be clinically significant and that the claim of equivalence can be misleading if a study has not been conducted at a sufficiently high level. In addition, physicians must be a bit skeptical of judgments that do not include the basic requirements of information, including the definition and justification of the equivalence margin, the calculation of the size of the sample bearing in mind this margin, the presentation of both analysis (intention-to-treat and by protocol), and provide confidence intervals for the results. Equivalence and inferiority studies are not indicated in certain areas. If one follows the required strict adherence to the specific methodology, such studies can provide new and important knowledge.

Keywords	Palabras clave
Clinical trial	Ensayo clínico
Randomized controlled trial	Ensayo controlado aleatorio
Therapeutic equivalency	Equivalencia terapéutica

El estudio clínico aleatorio (ECA) es generalmente aceptado como el mejor método para comparar los efectos de los tratamientos.[1,2] Muy a menudo el objetivo de un ECA es mostrar que un nuevo tratamiento es superior a una terapia o placebo establecido, es decir, que se planifica y se lleva a cabo como un estudio de superioridad. A veces el objetivo de un ECA se establece solo para demostrar que una nueva terapia no es superior, sino igual o no inferior a una terapia establecida, es decir, que los ECA se planifican y llevan a cabo como estudios de equivalencia o estudios de no inferioridad.[3] Dado que este tipo de pruebas tienen diferentes objetivos, difieren significativamente en diversos aspectos metodológicos.[4] La conciencia de las diferencias metodológicas generalmente es bastante limitada. Por ejemplo, es una creencia bastante común que el fracaso de encontrar una diferencia significativa entre los tratamientos en un estudio de superioridad implica que las terapias tienen el mismo efecto o son equivalentes.[5-10] Sin embargo, esta conclusión no es correcta debido a un considerable riesgo de pasar por alto un efecto clínicamente relevante debido al insuficiente tamaño de la muestra.

La Declaración CONSORT (Consolidated Standards of Reporting Trials), que incluye una lista y un diagrama de flujo, es una guía desarrollada para ayudar a los autores a mejorar la divulgación de los resultados de los ensayos controlados aleatorios. Se ha actualizado recientemente, en 2010.[1] Su enfoque principal es individualmente aleatorizado, con dos grupos paralelos que evalúan la posible superioridad de un tratamiento en comparación con el otro. La Declaración CONSORT se extendió a otros diseños de ensayo, como los ensayos aleatorizados en racimo, y a partir de esto se formularon recomendaciones para la equivalencia de ensayos y resultados en 2006. El propósito de este trabajo es revisar la metodología de los diferentes tipos de estudios, con especial referencia a las dife-

[a]Investigación en Salud
[b]Jefatura del Departamento de Hipertensión Pulmonar y Función Ventricular Derecha
[c]División de Enseñanza
[d]Dirección de Enseñanza e Investigación
[e]Dirección Médica
[f]Dirección General
[g]Academia Nacional de Medicina, A.C.

[a,b,c,d,e,f]Hospital de Cardiología, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social

Distrito Federal, México

Comunicación con: Martin Rosas-Peralta
Correo electrónico: mrosas_peralta@hotmail.com

Recibido: 02/03/2015 Aceptado: 04/03/2015

Resumen

Los médicos deben recordar siempre que un resultado negativo en una prueba de superioridad nunca probará que las terapias investigadas son equivalentes; más a menudo puede haber un gran riesgo de error de tipo 2 (resultado falso negativo). Los estudios de equivalencia y de no inferioridad exigen altos estándares para proporcionar resultados confiables. Los médicos deben tener en cuenta sobre todo que los márgenes de equivalencia suelen ser demasiado grandes como para ser clínicamente significativos y que la reclamación de esta puede ser engañosa si un estudio no se ha llevado a cabo a un nivel suficientemente elevado. Además, los médicos deben ser un poco escépticos de análisis que no incluyan los requisitos básicos de información, incluida la definición y la justificación del margen de equivalencia, el cálculo del tamaño de la muestra (deben tomar en cuenta este margen), la presentación de ambos tipos de análisis (por intención de tratar y por protocolo), así como los intervalos de confianza para los resultados. Los estudios de equivalencia y de no inferioridad se indican en ciertas áreas. Si se sigue la estricta adherencia necesaria a la metodología específica, tales estudios pueden proporcionar nuevos e importantes conocimientos.

rencias con respecto a la planificación, la ejecución, el análisis y la presentación de la prueba, así como la extensión de la declaración de CONSORT.[1] En este contexto se examinarán los conceptos estadísticos más relevantes. Algunos de los puntos importantes se ilustrarán con ejemplos.

Estudios de superioridad

La estimación del tamaño de la muestra y el poder de un ECA

Un aspecto importante en la planificación de cualquier ECA es estimar el número de pacientes necesarios, es decir, el tamaño de la muestra. En este sentido, los diversos tipos de estudios se diferencian.[1,2,11] Un estudio de superioridad pretende demostrar la superioridad de una nueva terapia en comparación con una terapia o placebo establecido. La siguiente descripción se aplica a un análisis de superioridad. Las características por las que una equivalencia o un estudio de no inferioridad se diferencian se describirán más adelante. Para estimar el tamaño de la muestra es necesario considerar algunos aspectos importantes, por ejemplo: ¿Por cuánto debe la nueva terapia ser mejor que la terapia de referencia? Este efecto adicional del tratamiento en cuestión frente al tratamiento de referencia se denomina *diferencia relevante* o al menos *de significado clínico*. A menudo se denota por la letra griega Δ (delta) (figura 1).

¿De cuánto sería la diferencia en el efecto entre los dos grupos al estar influenciada por factores aleatorios? Como cualquier otra medición de un efecto de tratamiento biológico, este estará sujeto a una variación considerable "al azar", que necesita ser determinado y tomado siempre en cuenta. La magnitud de la variación se describe en términos estadísticos por la desviación estándar S o por la varianza S^2 (figura 1 C), la cual tendría que ser obtenida a partir de un estudio piloto o de estudios similares previamente publicados.

La varianza del efecto de los tratamientos

Este estudio debe demostrar con la mayor precisión posible la verdadera diferencia entre el efecto de los tratamientos. Sin embargo, debido a la variación aleatoria el resultado final del estudio puede desviarse de la verdadera diferencia y dar resultados erróneos. Si, por ejemplo, la hipótesis nula H_0 de ninguna diferencia fuera cierta, podría darse incluso que el análisis en algunos casos sea mostrar una diferencia. Este tipo de error es llamado *error tipo 1* (*falso positivo*) (figura 1) y tendría la consecuencia de la introducción de una terapia ineficaz.

Si por el contrario la hipótesis alternativa H_Δ de la diferencia (Δ*: delta)* fuera verdad, el análisis podría, en algunos casos, no mostrar una diferencia significativa. Este tipo de error es el *error tipo 2* o *falso negativo* (figura 1), el cual tendría la consecuencia de rechazar una terapia eficaz. Así que uno tiene que especificar los riesgos de errores tipo 1 y tipo 2 que serían aceptables durante el análisis. Lo ideal sería que los riesgos del error tipo 1 y tipo 2 estuvieran cerca de cero, pero esto requeriría de estudios extremadamente grandes.

Muy a menudo el riesgo de cometer el error tipo 1 α se especifica al 5 %. En este trabajo, *alfa* significa que el riesgo de cometer el error tipo 1 está solo en una dirección, es decir, ya sea hacia arriba o hacia abajo de H_0; así, α = 5 %. Sin embargo, en muchas situaciones podría ser de interés la detección de los efectos tanto beneficiosos como perjudiciales de la nueva terapia en comparación con el tratamiento de control, es decir, uno estaría interesado en la prueba de "doble cara" o de "dos colas" de una diferencia en dirección

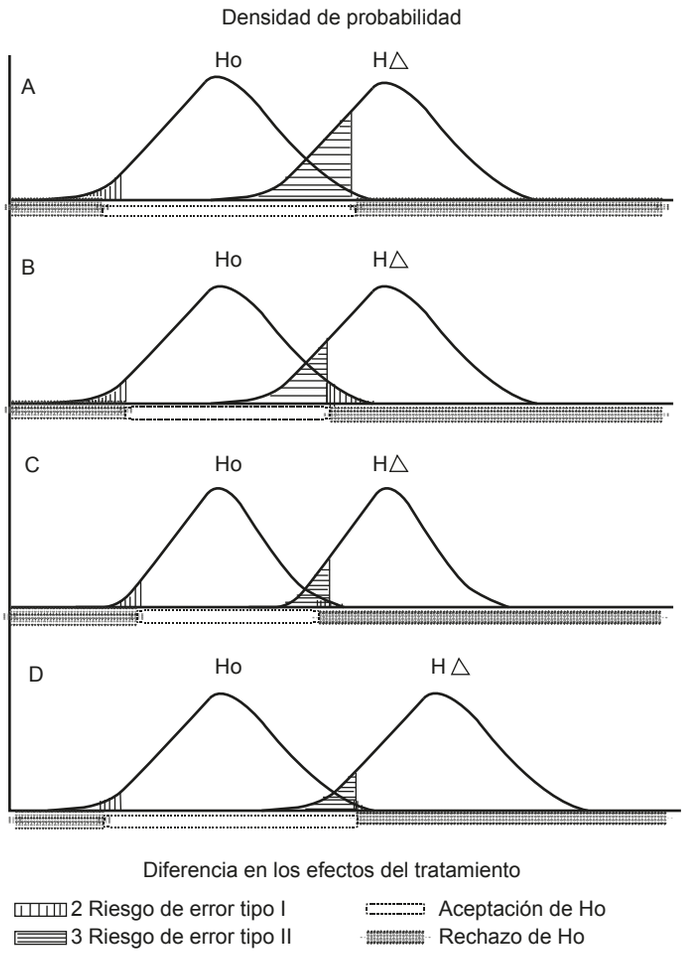

Figura 1 Ilustración de los factores que influyen en el tamaño de la muestra de un estudio. La diferencia del efecto encontrado en un estudio estará sujeta a una variación aleatoria. La variación se ilustra por las curvas normales de distribución en forma de campana para una diferencia de cero correspondiente a la hipótesis nula (H_0) y para una diferencia de Δ correspondiente a la hipótesis alternativa (H_Δ). Las áreas definidas bajo las curvas indican la probabilidad de una diferencia dada para que sea compatible con H_0 o H_Δ, respectivamente. Si la diferencia se encuentra cerca de H_0, se podría aceptar H_0. Cuanto sea más la diferencia de H_0, menos probable es que sea H_0. Si existe la probabilidad de que H_0 se vuelva muy pequeña (menor que la especificada error de tipo 1) al riesgo 2 α (siendo α, encontrada en la cola de la curva), se podría rechazar la H_0. Las curvas de distribución de la muestra disponen algunas coincidencias. Un gran traslape resultará en un considerable riesgo de error de interpretación; en particular, el riesgo de error de tipo 2 puede ser sustancial, como se indica en la figura. Una cuestión importante sería reducir el riesgo de error tipo 2 o β (y aumentar el poder de 1 - β) a un nivel razonable. Tres maneras de hacerlo se muestran en (B-D), y hay siempre una situación de referencia. (B) Aislados incrementos del 2 α disminuirá β y aumentará la potencia. A la inversa, una disminución aislada de 2α aumentará β y disminuirá la potencia. (C) El estrechamiento aislado de las curvas de distribución de la muestra (por el aumento de tamaño de la muestra 2 N o la disminución de la varianza de la diferencia S^2) disminuirá a β y aumentará la potencia. A la inversa, el ensanchamiento aislado de las curvas de distribución de la muestra (por la disminución de tamaño de la muestra o por el aumento de la varianza de la diferencia) aumentará a β y disminuirá la potencia. (D) El aumento aislado de Δ (efecto terapéutico mayor) disminuirá β y aumentará la potencia. A la inversa, con la disminución aislada de Δ (que hará el efecto terapéutico más pequeño) se incrementará β y disminuirá la potencia

tanto "hacia arriba" como "hacia abajo" (figura 1). Por lo tanto tendríamos un lugar específico para el riesgo de error tipo 1, o sea 2 α ($\alpha_{\text{hacia arriba}} + \alpha_{\text{hacia abajo}}$), es decir, 2 α = 5 %.

El error tipo 2 es el riesgo β (beta) y generalmente se especifica entre 10 y 20 % en los estudios clínicos. Puesto que un valor dado de Δ estará siempre por encima o por debajo de cero (H_0), el riesgo de cometer el error tipo 2 o beta, estará siempre de un solo lado. Así, cuanto menor sea el valor de β, mayor será la probabilidad de su complementaria o sea, $1 - \beta$ de aceptar H_Δ cuando de hecho sí es cierto.

A 1-β se le llama el poder o la potencia de la prueba, ya que establece la probabilidad de encontrar Δ si realmente existiera esta diferencia. De los valores dados a Δ, S^2, α y β, el número necesario (N) de los pacientes en cada grupo se puede estimar con el uso de la siguiente fórmula general que es relativamente simple:

$$N = (Z_{2x} + Z_\beta) \times S^2 / \Delta^2$$

Donde, $Z_{2\alpha}$ y Z_β son las desviaciones normalizadas correspondientes a los niveles de los valores definidos de 2 α (cuadro I, izquierda) y β (cuadro I, derecha), respectivamente. Si por alguna razón se quiere probar la diferencia en una sola dirección (pruebas de "una cola") se debe reemplazar $Z_{2\alpha}$ con Z_α en la fórmula y aplicar el lado derecho del cuadro I. La fórmula es aproximada, pero da en la mayoría de los casos una buena estimación del número necesario de pacientes. Para un estudio con dos grupos paralelos de igual dimensión, el tamaño total de la muestra será de 2 N.

Los valores utilizados para 2 α, β y Δ deben ser siempre decididos por el investigador, no por el estadístico. Los valores elegidos deben tener en cuenta la enfermedad, su etapa, la eficacia y los efectos secundarios de la terapia de control, así como una estimación de la cantidad de efecto adicional que se puede esperar razonablemente por la nueva terapia.

Si por ejemplo la enfermedad es bastante benigna con un pronóstico relativamente bueno y la nueva terapia es más cara y puede tener más efectos secundarios que un tratamiento de control bastante eficaz, se debe especificar un valor relativamente mayor de Δ y β y uno más pequeño para 2 α, debido a que la nueva terapia solo sería interesante si es notablemente mejor que el tratamiento de control.

Si por el contrario la enfermedad es agresiva, y la nueva terapia es más barata o puede tener menos efectos secundarios que la terapia de control (no muy eficaz), se debe intentar especificar un valor relativamente menor para Δ y β y uno más grande para 2 α, debido a que la nueva terapia se haría interesante, incluso si su valor es solo ligeramente mejor que el tratamiento de control.

Como se mencionó anteriormente, 2 α normalmente se especifica al 5 % o 0.05, pero uno puede justificarse a los valores de 0.10 o 0.01 en ciertas situaciones, como también se mencionó anteriormente. El valor de β normalmente se especifica entre 0.10 y 0.20, pero en situaciones especiales un mayor o menor valor puede estar justificado. El valor de Δ debe decidirse sobre bases clínicas y se traduce como la ganancia terapéutica relevante de la nueva terapia a partir de tomar en cuenta la prevalencia de la enfermedad y su pronóstico sobre la eficacia de la terapia de control y lo que razonablemente se puede esperar de la nueva terapia. Los datos preliminares de estudios experimentales o de los datos de observaciones históricas pueden ser pautas para la elección de la magnitud de Δ. Aunque suele ser tentador especificar una proporción de Δ relativamente grande para, por lo tanto, necesitar menos pacientes, Δ nunca se debe especificar más grande de lo que es biológicamente razonable o clínicamente trascendente. Así, siempre será poco ético realizar estudios con objetivos poco realistas. La figura 1 ilustra los efectos del riesgo de error β o tipo 2 y, por tanto, también de la potencia $(1 - \beta)$ de cambio de 2 α, N , S^2 y Δ . Así, β se reducirá y el poder $1 - \beta$ se incrementará si 2 α se incrementa (figura 1B), si el tamaño de la muestra se incrementa (figura 1C), y si Δ aumenta (figura 1D). El tamaño estimado de la muestra debe aumentar en proporción a la pérdida esperada de los pacientes durante el seguimiento debido a los abandonos y retiros.

El intervalo de confianza

Un concepto importante que indica la confianza del resultado obtenido en un ECA es la amplitud del *intervalo de confianza*, popularmente conocido con las siglas IC, de la diferencia *delta* en el efecto entre las terapias investigadas.[1,2]

Cuanto más estrecho sea el intervalo de confianza, más confiable será el resultado. En general, la anchura de este intervalo se determina por el tamaño de la muestra. Un gran tamaño de la muestra daría lugar a un estrecho intervalo de confianza. Normalmente, este se estima al 95 %. Así, este intervalo establece que en promedio incluirá la verdadera diferencia en 95 de cada 100 estudios similares. Esto se ilustra en la figura 2, en la que 100 muestras de prueba del mismo tamaño se han extraído al azar de la misma población. Es importante observar que en 5 de las 100 muestras del intervalo de confianza 95 % de la diferencia en el efecto D no se incluye la verdadera diferencia encontrada en la población, es decir se acepta un error de 5 % o 0.05.

Cuando los intervalos de confianza están alineados de acuerdo con su promedio (figura 2C), la variación en relación con el valor real en la población se vuelve aún más clara. Si la simulación se lleva a cabo a una escala aún mayor, la distribución de probabilidad de la verdadera diferencia en la población, dados los resultados de una determinada muestra de estudio, se siguen en una distribución normal, como se muestra en la figura 3.[2]

Se ve que la probabilidad de que la diferencia real en la población sea máxima en la diferencia D encontrada en la muestra y que disminuye con valores más altos y más bajos. La figura también ilustra el intervalo de confianza del 95 %, que es el intervalo que incluye el 95 % del promedio del área total bajo la curva de probabilidad normal. Esta área se puede calcular a partir de la diferencia D y su error estándar (ESD). Para hacer más seguro que la verdadera diferencia esté incluida en

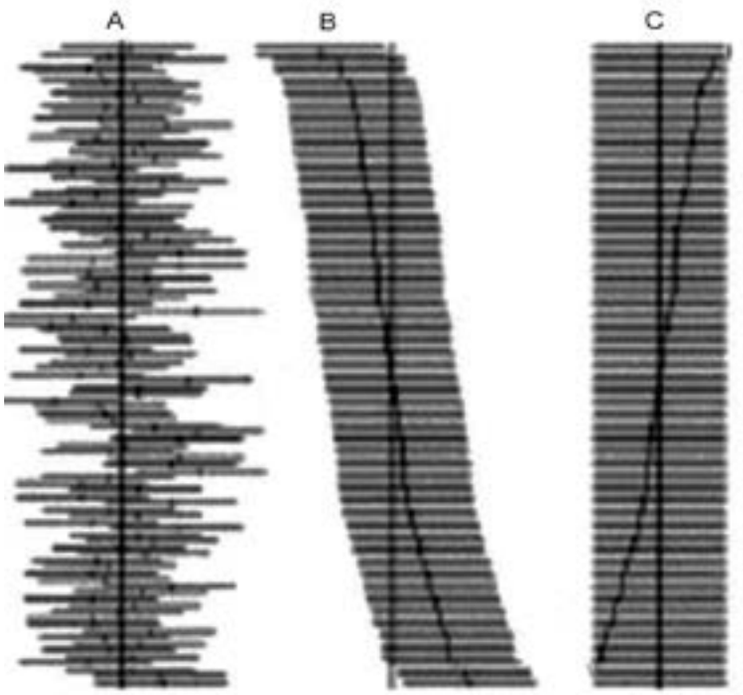

Figura 2 Ilustración de la variación de los límites de confianza en muestras aleatorias (simulación por ordenador). (A) noventa y cinco por ciento de intervalos de confianza en 100 muestras aleatorias del mismo tamaño de la misma población alineadas según el valor real en la población. En cinco de las muestras del intervalo de confianza del 95 % no se incluye el valor real que se encuentra en la población. (B) Los mismos intervalos de confianza están aquí ordenados de acuerdo con sus valores. (C) Cuando los intervalos de confianza se alinean según su media, su variación en relación con el valor real de la población se ve claramente de nuevo. Esta presentación corresponde a cómo los investigadores iban a ver el mundo. Investigan las muestras con el fin de extrapolar los resultados a la población. Sin embargo, la imprecisión potencial de extrapolar a partir de una muestra a la población es evidente (especialmente si el intervalo de confianza es amplio). Por lo tanto, es importante mantener los intervalos de confianza más bien estrechos. Esto significa realizar estudios relativamente más grandes.

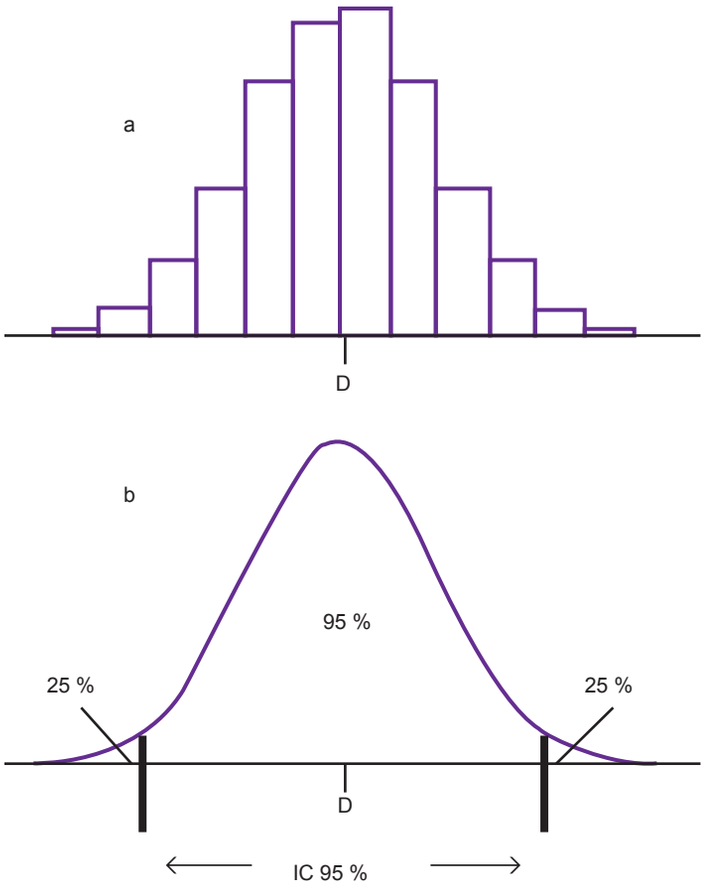

Figura 3 (A) Histograma que muestra la distribución de la verdadera diferencia en la población en relación con la diferencia D encontró en la muestra de estudio (simulación por ordenador de 10 000 muestras). (B) La curva de probabilidad de distribución normal de la verdadera diferencia en la población en relación con la diferencia D se encontró en una muestra de prueba. Se muestra el intervalo de confianza del 95 % (IC 95 %)

el intervalo de confianza, se puede calcular uno al 99 %, lo que sería más amplio, ya que debe incluir el medio 99 % de la superficie total de probabilidad.

El riesgo de error tipo 2 (o β) de pasar por alto una diferencia (Δ) real

Si el intervalo de confianza del 95 % de Δ incluye al cero, entonces no hay diferencia significativa en el efecto entre dos terapias. Sin embargo, esto no quiere decir que se puede concluir que los efectos de las terapias son los mismos. Todavía puede haber una verdadera diferencia en el efecto entre las terapias que los ECA no hayan sido capaces de detectar, debido al tamaño insuficiente de la muestra o al poder.

El riesgo de pasar por alto una cierta diferencia en el efecto de Δ entre las terapias es el riesgo de cometer el error de tipo 2 o β. En algunos casos, este riesgo puede ser sustancial. Un ejemplo de este patrón se presenta en el caso 1.

Caso 1

Se conoce que los casos no tratados previamente de miocarditis por virus de Coxackie B/genotipo 1 al usar interferón y ribavirina durante tres meses inducen respuesta virológica sostenida en aproximadamente 40 %. Uno desea probar si un nuevo régimen terapéutico puede aumentar la respuesta sostenida en este tipo de pacientes a 60 % con una potencia ($1 - \beta$) de 80 %. El tipo 1 de riesgo de error ($2\ \alpha$) debe ser 5 %. Hay que calcular el número de pacientes necesarios para este estudio. Para la comparación de proporciones, como en este estudio, la varianza de la diferencia (S^2) es igual a $p_1 (1 - p_1) + p_2 (1 - p_2)$, donde p_1 y p_2 son las proporciones con respuesta en los grupos de comparación. Así, tenemos:

$2\alpha = 0.05 \rightarrow Z_{2\alpha} = 1.96$ $\beta = 0.20 \rightarrow Z_\beta = 0.84$

$p_1 = 0.4$ $p_2 = 0.6$ $\Delta = 0.2$

Con el uso de $N = (Z_{2\alpha} + Z_\beta)^2 \times p_1 (1 - p_1) + p_2 (1 - p_2) / \Delta^2$ uno obtiene:

$N = (1.96 + 0.84)^2 \text{ X } (0.4 \text{ X } 0.6 + 0.6 \text{ X } 0.4) / 0.2^2$

$= 7.84 \text{ X } 0.48 / 0.04 = 94$

Por lo tanto, el número necesario de pacientes ($2N$) sería 188.

Sin embargo, debido a diversas dificultades solo 120 pacientes (60 en cada grupo) de este tipo podrían ser reclutados. Al resolver la fórmula general el tamaño de la muestra de acuerdo con Z_β se obtiene:

$$Z_\beta = \frac{\sqrt{N}}{S} \text{ x } \Delta - Z_{2\alpha}$$

Al utilizar esta fórmula, el poder de la prueba con el número reducido de pacientes puede ser estimado como sigue:

$Z_\beta = \frac{\sqrt{60}}{\sqrt{0.48}} \text{ x } 0.2 - Z_{2\alpha}$ $Z_\beta = 7.75 / 0.69 \text{ X } 0.2 - 1.96$

$= 0.29$

Si se usa la parte derecha del cuadro I con interpolación β se convierte 0.39. Así, con este número limitado de pacientes, el poder $1 - \beta$ es ahora solo 0.61 o 61 % (una reducción clara en vez del tradicional 80 %). Este poder marcadamente reducido disminuye seriamente las posibilidades de demostrar un efecto significativo del tratamiento. Un cálculo *post hoc* del poder como este solo se puede utilizar para explicar por qué un análisis de superioridad no

es concluyente nos lleva a recordar que *nunca puede ser utilizado esto* para apoyar el resultado negativo de una prueba de superioridad.

El resultado del estudio fue como sigue: la respuesta virológica sostenida se encontró en 26 de 60 (0.43 o 43 %) en el grupo control y en 35 de 60 (0.58 o 58 %) en el nuevo grupo de terapia. La diferencia *D* es 0.15 o 15 %, pero no es estadísticamente significativa ($p > 0.10$). Una fórmula aproximada simple para el error estándar de la diferencia es:

$$SED = \sqrt{p_1 (1 - p_1) / n_1 + p_2 (1 - p_2) / n_2}$$

$$= \sqrt{0.43 \times 0.57 / 60 + 0.58 \times 0.42 / 60} = 0.09$$

El intervalo de confianza del 95 % para *D* es $D \pm Z_{2\alpha} \times \text{SED} = 0.15 \pm 1.96 \times 0.09$ o −0.026 − 0.326 (−2.6 % a 32.6 %), lo cual es bastante amplio, ya que incluye tanto el cero como a *Δ*. El riesgo de error tipo 2 a un efecto de 20 % (correspondiente a *Δ*) se puede estimar de la siguiente manera:

$$Z_\beta = (\Delta - D) / \text{SED} = (0.20 \text{ a } 0.15) / 0.09 = 0.55$$

Si se usa la parte derecha del cuadro I con interpolación *β* se convierte 0.29. Así, el riesgo de haber pasado por alto un efecto de 20 % es del 29 %. Esto es una consecuencia del número pequeño de pacientes incluidos y la potencia reducida del estudio. La situación corresponde a la que se ilustra en la figura 4. Como se ve en esta figura el resultado de un ECA negativo como este no descarta que la verdadera diferencia puede ser *Δ* , ya que el riesgo de error tipo 2 (o *β)* de pasar por alto un efecto de *Δ* es sustancial.

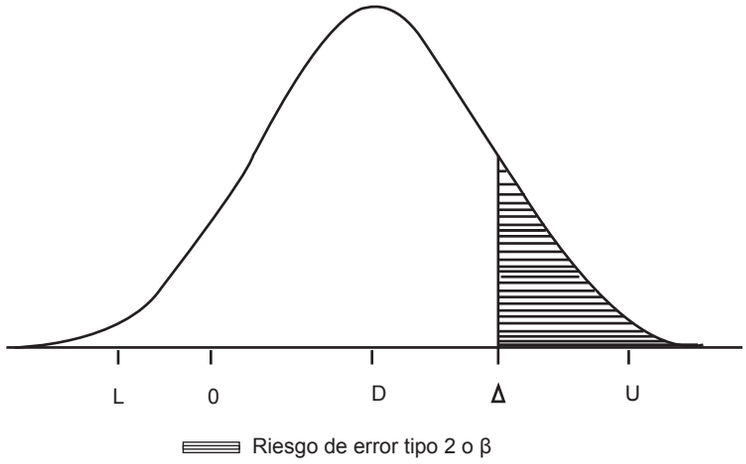

Figura 4 Ilustración del riesgo de error β o tipo 2 (área en gris) en un ECA que muestra una diferencia D en efecto, que no es significativo, ya que el cero (0) de diferencia se encuentra en el límite bajo (L) y el límite alto (U) del intervalo de confianza al 95 %. El riesgo de error tipo 2 que tiene el valor de Δ es sustancial

Estudios de equivalencia

El propósito de un estudio de equivalencia es establecer efectos idénticos de los tratamientos que se comparan.[12-17] Tener efectos equivalentes significaría tener un valor de *Δ* de cero. Como se ve a partir de la fórmula para la estimación del tamaño de muestra (ver arriba) esta división significaría cero, lo que no es posible. Dividiendo por un muy pequeño valor *Δ* se traduciría en una muestra de gran tamaño, pero realista. Por lo tanto, como un compromiso manejable, el objetivo de un estudio de equivalencia sería determinar si la diferencia de efectos entre dos terapias se encuentra dentro de un intervalo pequeño de $-\Delta$ a $+\Delta$.

Un estudio de equivalencia sería relevante si la nueva terapia fuera más simple, estuviera asociada a menos efectos secundarios o menos caros, e incluso si no se esperara que tuviera un efecto terapéutico mayor que el tratamiento de control. Lo anterior es crucial para especificar un tamaño correspondiente de *Δ*.[14,17]

Esto no es simple. Uno debe estar dispuesto a limitar lo más posible la aceptación de una nueva terapia, que es inferior a la terapia de control. Por lo tanto, el valor de *Δ* debe especificarse como pequeño y en cualquier caso menor que el valor más pequeño que pudiera representar una diferencia clínicamente significativa. Como regla general, *Δ* debe especificarse a no más de la mitad del valor que se puede utilizar en un estudio de superioridad.[13] La equivalencia entre las terapias se demostraría si el intervalo de confianza para la diferencia en el efecto entre las terapias resulta que se sitúe totalmente entre $-\Delta$ y $+\Delta$.[13] La figura 5 ilustra las conclusiones que se pueden extraer de la posición de los límites de confianza para la diferencia en el efecto encontrado en el estudio realizado.

Es crucial comprender que en el estudio de equivalencia se invierten los roles de las hipótesis nula y alternativa, es decir, la hipótesis nula relevante es que una diferencia de al menos *Δ* existe, y el objetivo del estudio es refutar esto en favor de la hipótesis alternativa de que no existe diferencia.[13] Aunque esta situación es un reflejo de la superioridad de análisis similar, resulta que el método para la estimación del tamaño de la muestra también es similar en los dos tipos de análisis, aunque *Δ* tiene diferentes significados en los estudios de superioridad y de equivalencia.

Caso 2

En los mismos pacientes que se describieron en el caso 1 se desea comparar en un ECA el supuesto de

equivalencia terapéutica del régimen actual de interferón y ribavirina (con una respuesta sostenida de 40 %) con otro nuevo régimen terapéutico de bajo costo que tiene menos efectos secundarios.

Hay que calcular el número de pacientes necesarios para este estudio. La potencia ($1 - \beta$) de la prueba debe ser del 80 %. El riesgo de error tipo 1 de ($2\ \alpha$) debe ser del 5 %. Las terapias se consideran equivalentes si el intervalo de confianza para la diferencia en proporción con la respuesta sostenida cae enteramente dentro del intervalo de ± 0.10 % o ± 10 %. Por lo tanto, Δ se especifica a 0.10. Así tenemos:

$2\alpha = 0.05 \rightarrow Z_{2\alpha} = 1.96 \qquad \beta = 0.20 \rightarrow Z_{\beta} = 0.84$

$p_1 = 0.4 \quad p_2 = 0.4 \quad \Delta = 0.10$

Si usamos la misma expresión para la varianza de la diferencia (S^2), como en el caso 1, se obtiene el siguiente resultado:

$$N = (1.96 + 0.84)^2 \text{ X } (0.4 \text{ X } 0.6 + 0.4 \text{ X } 0.6) / 0.1^2$$

$$= 7.84 \text{ X } 0.48 / 0.01 = 376$$

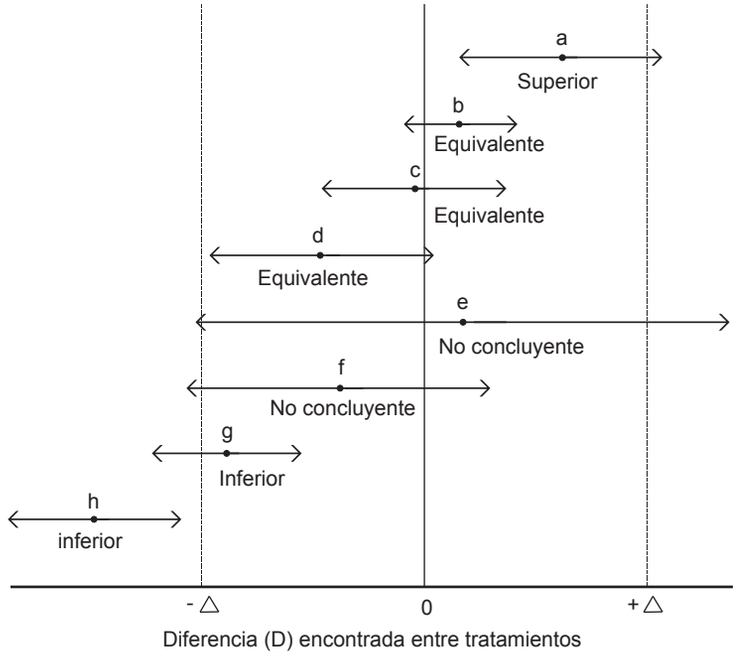

Figura 5 Los ejemplos de las diferencias de tratamiento observados (nueva terapia-terapia de control) con intervalos de confianza al 95 % y las conclusiones que se pueden extraer. (a) La nueva terapia es significativamente mejor que el tratamiento de control. Sin embargo, la magnitud del efecto puede ser clínicamente importante. (b-d) Las terapias pueden ser consideradas de efecto equivalente. (e-f) El resultado no es concluyente. (g) La nueva terapia es significativamente peor que el tratamiento de control, pero la magnitud de la diferencia puede ser clínicamente importante. (h) La nueva terapia es significativamente peor que el tratamiento de control

Por lo tanto, el número necesario de pacientes ($2\ N$) sería 752.

El estudio se llevó a cabo y el resultado demostró que se encontró respuesta virológica sostenida en 145 de 372 (0.39 o 39 %) en el grupo control y en 156 de 380 (0.41 o 41 %) en el nuevo grupo de terapia. La diferencia D fue de 0.02 o 2 %, pero no fue estadísticamente significativa ($p > 0.50$). El error estándar de la diferencia fue:

$$SED = \sqrt{p_1 (1 - p_1) / n_1 + p_2 (1 - p_2) / n_2}$$

El intervalo de confianza del 95 % para D fue $D \pm Z_{2\alpha} \times \text{SED} = 0.02 \pm 1.96 \times 0.036$ o −0.050-0.091 (−5.0 a 9.1 %). Dado que este intervalo de confianza se encontró completamente dentro del intervalo especificado para Δ (de −0.1 a 0.1), los efectos de las dos terapias se consideraron equivalentes. La situación corresponde a B o C en la figura 5.

Como en este ejemplo, el tamaño de la muestra necesario en un estudio de equivalencia a menudo será de al menos cuatro veces el de un estudio correspondiente de superioridad. Por lo tanto, los recursos necesarios serán más grandes.

Estudios de no inferioridad

El estudio de no inferioridad, que se relaciona con el análisis de equivalencia, no tiene como objetivo mostrar la equivalencia, sino solo mostrar que la nueva terapia no es peor que la terapia de referencia. Así, el estudio de no inferioridad está diseñado para demostrar que la diferencia en el efecto (nueva terapia frente a terapia de control) debe ser inferior a $-\Delta$. La no inferioridad de la nueva terapia sería entonces demostrada si el límite inferior de confianza para la diferencia en el efecto entre las terapias resulta estar por encima de $-\Delta$. La posición del límite superior de confianza no es de interés primordial. Así, el estudio de no inferioridad está diseñado como un estudio de un solo lado. Por esa razón el número necesario de pacientes siempre será menor que para un estudio de equivalencia correspondiente, como se ilustra en el siguiente caso.

Caso 3

Queremos llevar a cabo el estudio descrito en el caso 2, no como un estudio de equivalencia, sino como un estudio de no inferioridad. Así, el estudio debe ser de un solo lado en vez del estudio de equivalencia de dos colas. La única diferencia sería que

uno debe usar Z_{α} en lugar de $Z_{2\alpha}$. Para $\alpha = 0.05$ se obtiene $Z_{\alpha} = 1.64$ (lado derecho del cuadro I). De este modo se obtiene:

$$N = (1.64 + 0.84)^2 \text{ X } (0.4 \text{ X } 0.6 + 0.4 \text{ X } 0.6) / 0.1^2$$

$$= 6.15 \text{ X } 0.48 / 0.01 = 295$$

Por lo tanto el número necesario de pacientes ($2N$) sería 590.

El estudio se llevó a cabo y el resultado mostró que se encontró una respuesta virológica sostenida en 114 de 292 (0.39 o 39 %) en el grupo control y en 125 de 298 (0.42 o 42 %) en el grupo de terapia nueva. La diferencia D es 0.03 o 3 %, pero no es estadísticamente significativa ($p > 0.50$). El error estándar de la diferencia es:

$$SED = \sqrt{p_1 (1 - p_1) / n_1 + p_2 (1 - p_2) / n_2}$$

$$= \sqrt{0.39 \text{ X } 0.61 / 292 + 0.42 \text{ X } 0.58 / 298} = 0.040$$

El límite inferior de confianza unilateral del 95 % sería $D - Z_{\alpha} \times \text{SED} = 0.03 - 1.64 \times 0.040 = -0.036$ (−3.6 %). Dado que el límite inferior de confianza está por encima del límite especificado para Δ de −0.1, el efecto de la nueva terapia no es inferior a la terapia de control. Si el intervalo de confianza de dos colas para el 95 % está siendo estimado (el cual es recomendado por algunos incluso para el estudio de no inferioridad),[18] se obtiene $D \pm Z_{2\alpha} \times \text{SED} = 0.03 \pm 1.96 \times 0.040$ o −0.048-0.108 (de −4.8 % a 10.8 %). El límite inferior de confianza aún se encuentra por encima de −0.1, pero el límite superior de confianza está por encima de 0.1 (el límite superior de la equivalencia; véase el caso 2). Por lo tanto, la nueva terapia puede ser ligeramente mejor que el tratamiento de control. El riesgo de error tipo 2 es alto y si tiene un efecto de 0.1 o 10 % podría ser estimado como sigue: $Z_{\beta} = (\Delta - \text{D}) / \text{SED} = (0.10 - 0.03) / 0.04 = 1.75$. Si se usa la parte derecha del cuadro I con interpolación de *β, se* convierte 0.04, es decir, más bien se trata de un riesgo pequeño.

Otros factores

Dado que el objetivo de un estudio de equivalencia o de un estudio de no inferioridad es diferente, no es el mismo incentivo para eliminar factores que pudieran oscurecer cualquier diferencia entre los tratamientos, como es el caso de un estudio de superioridad. Así, en algunos casos el hallazgo de equivalencia puede ser debido a las deficiencias de los estudios, como el tamaño pequeño de la muestra, la falta de doble ciego, la falta de asignación aleatoria oculta, las dosis incorrectas de los fármacos, los efectos de la medicina concomitante o la recuperación espontánea de los pacientes sin la intervención médica.[19]

Tanto el estudio de equivalencia como el estudio de no inferioridad deben reflejar lo más fielmente posible los métodos utilizados en los estudios de superioridad que anteriormente evaluaron el efecto de la terapia de control frente al placebo. En particular, es importante que los criterios de inclusión y exclusión que definen la población de pacientes, el cegamiento, la asignación al azar, el esquema de dosificación del tratamiento estándar, el uso de medicación concomitante (y otras intervenciones), la variable de respuesta primaria y el calendario de sus mediciones sean los mismos que en los estudios anteriores de superioridad, los cuales evaluan la terapia de referencia que se utiliza en la comparación. Además, se debe prestar atención a la conformidad del paciente, a la respuesta durante cualquier plazo de tiempo y a la magnitud de las pérdidas de pacientes (y las razones de estas). Estos estudios no deben ser diferentes de estudios anteriores de superioridad.

Análisis por intención de tratar y por protocolo

Un punto importante en el análisis de los estudios de equivalencia y de no inferioridad estriba en la disyuntiva de utilizar un *análisis por intención de tratar* o un *análisis por protocolo*. En un estudio de superioridad, en el que el objetivo es decidir si dos tratamientos son diferentes, el análisis por intención de tratar es generalmente conservador: la inclusión de los violadores de protocolo y los retiros por lo general tienden a hacer que los resultados de los dos grupos de tratamiento sean más similares. Sin embargo, para un estudio de equivalencia o no inferioridad este efecto ya no es conservador: cualquier eliminación de la diferencia entre los grupos de tratamiento aumentará la probabilidad de encontrar la equivalencia o no inferioridad.

Un análisis por protocolo compara pacientes según el tratamiento realmente recibido e incluye solo a aquellos que cumplieron con los criterios de entrada y siguieron correctamente el protocolo. En un estudio de superioridad este enfoque puede tender a mejorar cualquier diferencia entre los tratamientos en lugar de disminuir, dado que el ruido no informa acerca de los pacientes que se retiran. En un estudio de equivalencia o no inferioridad ambos tipos de análisis se deben realizar y la equivalencia o no inferioridad solo pueden establecerse si ambos análisis lo apoyan. Para garantizar la mejor calidad posible del análisis, es importante recopilar datos de seguimiento completo de todos los pacientes asignados tanto al azar como por protocolo,

Cuadro I Tabla abreviada de la distribución normal estandarizada (adaptada para este trabajo) (ver nota abajo)

Probabilidad a doble cola		Una cola de probabilidad			
					
$Z_{2\alpha}$	2α	Z_α o Z_β	α o β	Z_α o Z_β	α o β
3.09	0.002	3.09	0.001	-0.25	0.60
2.58	0.01	2.58	0.005	-0.39	0.65
2.33	0.02	2.33	0.010	-0.52	0,70
1.96	0.05	1.96	0.025	-0.67	0.75
1.64	0,1	1.64	0.05	-0.84	0.80
1.28	0,2	1.28	0.10	-1.04	0.85
1.04	0.3	1.04	0.15	-1.28	0.90
0.84	0.4	0.84	0.20	-1.64	0.95
0.67	0,5	0.67	0.25	-1.96	0.975
0.52	0,6	0.52	0.30	-2.33	0.990
0.39	0,7	0.39	0.35	-2.58	0.995
0.25	0.8	0.25	0.40	-3.09	0.999
0.13	0.9	0.13	0.45	-3.29	0.9995
0.00	1.0	0.00	0.50	-3.72	0.9999

El área total bajo la curva de distribución normal es una. El área bajo una parte dada de la curva da la probabilidad de una observación de estar en esa parte. En el eje de x se indica la «densidad de probabilidad», que es más alto en el centro de la curva y disminuye en cualquier dirección hacia las colas de la curva. La distribución normal es simétrica, es decir, la probabilidad de *Z* a infinito más (lado derecho del cuadro) es el mismo que el de - *Z* a -∞. El lado derecho de la mesa da la probabilidad de un solo lado de una determinada *Z* -valor de la *x* eje x a + ∞. El lado izquierdo de la tabla da la probabilidad de dos caras como la suma de la probabilidad de una Z-valor positivo dado a + ∞ y la probabilidad a partir del correspondiente negativo *Z* -valor a -∞

independientemente de si se encuentran luego de tener criterios de ingreso fallidos, si dejan de usar la medicación del estudio antes de tiempo, o si violan el protocolo de algún otro modo.[20] Un enfoque tan rígido para la recopilación de datos permite la máxima flexibilidad durante su posterior análisis y por lo tanto proporciona una base más sólida para las decisiones.

El problema más común en los estudios de equivalencia o no inferioridad reportados es que se planifican y se analizan como si fueran estudios de superioridad y que la falta de una diferencia estadísticamente significativa se toma como prueba de equivalencia.[7,8,9,10] Por lo tanto, parece que existe la necesidad de un mejor conocimiento de cómo se deben planificar los estudios de equivalencia y de no inferioridad, y de cómo se deben realizar sus análisis e informes.

Garantizar una alta calidad

Un estudio reciente informó sobre la calidad de los estudios de equivalencia publicados.[21] Uno de los hallazgos fue que algunos estudios habían sido planeados como estudios de superioridad, pero se presentaron como si hubieran sido estudios de equivalencia después del fracaso de demostrar la superioridad, ya que no incluyeron un margen de equivalencia. En ese estudio también se afirma que un tercio de los informes que incluyeron un cálculo del tamaño de la muestra habían omitido elementos necesarios para reproducirlo; un tercio de los informes describió un intervalo de confianza cuyo tamaño no estaba en conformidad con el riesgo de error tipo 1 utilizado en el cálculo del tamaño de la muestra; y la mitad de los informes que

utilizaron pruebas estadísticas no tomaron los márgenes en cuenta. Además, solo el 20 % de los estudios encuestados proporcionó los cuatro requisitos básicos necesarios: el margen de equivalencia definida, el cálculo del tamaño de la muestra con base en ese margen, el análisis del protocolo por intención de tratar y el intervalo de confianza para el resultado. Solo el 4 % de los estudios dio una justificación, que es esencial, para el margen utilizado.

Una extensión en relación con los estudios de equivalencia y de no inferioridad es la declaración CONSORT sobre publicaciones de ECA.[1,18,22,23,24] Esto incluye la descripción de las razones para la adopción de un diseño de equivalencia o no inferioridad, cómo se incorporaron las hipótesis de estudio en el diseño, la elección de los participantes, las intervenciones (especialmente el tratamiento de referencia) y los resultados derivados de los métodos estadísticos, incluido el cálculo del tamaño de la muestra y la forma en la que el diseño afecta a la interpretación y las conclusiones.[18]

Declaración de conflicto de interés: los autores han completado y enviado la forma traducida al español de la declaración de conflictos potenciales de interés del Comité Internacional de Editores de Revistas Médicas, y no fue reportado alguno en relación con este artículo.

Referencias

1. Piaggio G, Elbourne DR, Pocock SJ, Stephen JWE, Altman DG, for the CONSORT Group. Reporting of Noninferiority and Equivalence Randomized Trials Extension of the CONSORT 2010 Statement JAMA. 2012;308(24):2594-604.
2. Armitage P, Berry G. Statistical methods in medical research. 3rd ed. Oxford: Blackwell; 1994.
3. Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep. 1978;62:1037-40.
4. Fleiss JL. General design issues in efficacy, equivalency and superiority trials. J Periodontal Res. 1992;27:306-13.
5. Garrett AD. Therapeutic equivalence: fallacies and falsification. Stat Med. 2003;22:741-62.
6. Blackwelder WC. Proving the null hypothesis in clinical trials. Control Clin Trials. 1982;3:345-53.
7. 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.
8. Costa LJ, Xavier ACG, del Giglio A. Negative results in cáncer clinical trials – equivalence or poor accrual?. Control Clin Trials. 2004;25:525-33.
9. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: equivalency or error? Arch Surg. 2001;136:796-800.
10. Detsky AS, Sackett DL. When was a negative clinical trial big enough? How many patients you needed depends on what you found. Arch Intern Med. 1985;145:709-12.
11. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA. 1994;272:122-4.
12. Djulbegovic B, Clarke M. Scientific and ethical issues in equivalence trials. JAMA. 2001;285:1206-8.
13. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ. 1996;313:36-9.
14. Lange S, Freitag G. Choice of delta: requirements and reality results of a systematic review. Biomed J. 2005;47:12-27.
15. Durrleman S, Simon R. Planning and monitoring of equivalence studies. Biometrics. 1990;46:329-36.
16. Ebbutt AF, Frith L. Practical issues in equivalence trials. Stat Med. 1998;17:1691-701.
17. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. Control Clin Trials. 2002;23:2-14.
18. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW. CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA. 2006;295:1152-160.
19. Chan A-W, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA. 2004;291:2457-65.
20. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ. 2001;323:42-6.
21. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295:1147-51.
22. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA. 1996;276:637-9.
23. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet. 2001;357:1191-4.
24. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med. 2001;134:663-94.