Ronda clínica y epidemiológica
Estudios de superioridad vs. Estudios de no inferioridad

Carolina Estrada Pérez[1], Fabián Jaimes Barragán[2]

Las preguntas de investigación se pueden agrupar en categorías, según que intenten resolver dudas acerca del desempeño de una intervención terapéutica, una medida preventiva o una prueba diagnóstica, o de establecer el pronóstico de una enfermedad, entre otras. Respecto a la evaluación de una intervención, los diseños más comunes y ampliamente aceptados como el mejor método para establecer la eficacia terapéutica son los ensayos clínicos aleatorios (ECA), que tienen usualmente como objetivo demostrar que un nuevo tratamiento es superior al establecido (o al placebo, cuando aún no hay una terapia estándar para la enfermedad). Estos ECA, que a su vez son los más comunes en términos terapéuticos, se llaman estudios de superioridad. Sin embargo, algunas veces existe un imposible ético para hacer comparaciones con placebo, o no se pretende reemplazar el tratamiento estándar por una intervención de eficacia superior sino por una de eficacia similar, pero con menos efectos adversos, mayor facilidad de administración o menores costos. En estos casos, los ECA no quieren mostrar que la nueva terapia sea superior sino que es equivalente o no inferior a la terapia ya establecida, y estos son los llamados estudios de no inferioridad o de equivalencia. Aunque ambos están dentro del grupo de los ECA y tienen principios metodológicos similares, tienen particularidades en su diseño y análisis, concretamente en cuanto al cálculo del tamaño de la muestra y a la interpretación de los resultados. Como principio y requisito fundamental, cuando no se encuentran diferencias en el efecto de las intervenciones en un estudio de superioridad no se puede afirmar que dichas intervenciones sean equivalentes; y de manera similar, en un estudio de no inferioridad, el hecho de encontrar diferencias en la eficacia de las intervenciones no necesariamente demuestra que una sea superior a la otra.

Otra consideración fundamental en el diseño de este tipo de ECA está relacionada con la selección de la población de estudio. Dicha población de pacientes debe reflejar de la manera más fidedigna posible la población original en la que se demostró la eficacia del tratamiento estándar, también denominado control activo, dado que cualquier cambio en sus características puede a su vez trasladarse a una disminución o aumento del efecto establecido para la comparación (1).

[1] Estudiante de Medicina, Universidad de Antioquia, Medellín, Colombia.

[2] Profesor Titular, Grupo Académico de Epidemiología Clínica (GRAEPIC), Departamento de Medicina Interna, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia. Investigador, Unidad de Investigaciones, Hospital Pablo Tobón Uribe, Medellín, Colombia.

Financiación: trabajo apoyado parcialmente por la Estrategia de Sostenibilidad de la Universidad de Antioquia 2013-2014.

Correspondencia: Fabián Jaimes Barragán; fjaimes@udea.edu.co

Recibido: septiembre 04 de 2013
Aceptado: septiembre 26 de 2013

Tamaño de muestra en estudios de superioridad

Para estimar el tamaño de la muestra necesario (*N*) para obtener resultados precisos en un ECA es preciso considerar varios aspectos:

1. ¿Cuál es la diferencia en el efecto o diferencia clínicamente relevante que se espera encontrar (Δ)?
2. ¿Cuál es la variabilidad o la dispersión de esa diferencia (S^2)?
3. ¿Cuál es la cantidad de error aleatorio de tipo I ($Z_{2\alpha}$) y de tipo II (Z_β) que se puede tolerar?

Las anteriores consideraciones se resumen en la fórmula general de tamaño de muestra:

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

De esta estructura es posible entender cuáles características tienen una relación directamente proporcional (la variabilidad) o inversamente proporcional (el error aleatorio y la diferencia por detectar) con el tamaño de muestra requerido. Si para confirmar que una nueva terapia es mejor que el control la diferencia que se espera encontrar (Δ) es de gran magnitud, entonces se necesita un grupo pequeño de pacientes para demostrarlo. Por el contrario, si la eficacia o diferencia clínica esperada es mínima es necesario contar con un mayor tamaño de muestra para detectar esa diferencia. Por otra parte, una mayor variabilidad o dispersión de los datos (la denominada varianza: S^2) impide apreciar cualquier diferencia y por tanto obliga a un mayor número de pacientes.

La cantidad de error aleatorio tolerable está definida por dos probabilidades: la de encontrar por azar una diferencia inexistente, denominada error tipo I o alfa, y la de no encontrar una diferencia real, conocida como error de tipo II o beta. El error de tipo I puede ocurrir en dos direcciones (es decir, se puede encontrar una falsa diferencia a favor o en contra de la nueva terapia), por lo que en la terminología del cálculo de tamaño de muestra se habla de error alfa de dos colas o dos lados. Históricamente, aunque de manera un poco arbitraria, la comunidad científica ha considerado aceptable una probabilidad de error de tipo I de 5% (o 1 en 20, o 0,05) y una probabilidad de error de tipo II entre 10% y 20% (o 0,1-0,2, o 1 en 10-1 en 5). Estas probabilidades, para efectos de su adecuada inserción en la fórmula de tamaño de muestra, se transcriben con sus valores correspondientes en la variable continua "Z" de la distribución normal: $Z_{2\alpha}$ y Z_β, los cuales son mayores en las áreas de menor probabilidad. Es decir, una probabilidad de error I o alfa de 0,05, o de 0,025 por cada una de las dos posibles colas o direcciones, equivale a un valor "Z" de 1,96, y una probabilidad de error II o beta de 0,2 equivale a un valor "Z" de 0,84. Es decir, que aceptar una menor probabilidad de error aleatorio de cualquier tipo implica también un mayor valor de "Z" en el numerador de la fórmula, y por tanto un mayor tamaño de muestra requerido.

Tamaño de muestra en estudios de no inferioridad o equivalencia

El propósito de un estudio de equivalencia es establecer que el efecto de dos terapias es idéntico, es decir, que la equivalencia completa de dos tratamientos correspondería a una diferencia clínica (Δ) de cero. De la fórmula explicada anteriormente, que se aplica exactamente igual en este tipo de diseños, se observa que una división por cero es imposible, y que un valor extremadamente pequeño de Δ llevaría a tamaños de muestra irrealizables en estudios clínicos. Por consiguiente, como un compromiso razonable, el objeto de un estudio de equivalencia es determinar si la diferencia de efecto entre dos tratamientos está dentro de un pequeño intervalo estimado entre $-\Delta$ y $+\Delta$. Es crucial, por lo tanto, especificar un tamaño relevante de Δ que evite la aceptación de una nueva terapia que sea inferior al tratamiento estándar. Consecuentemente, la diferencia Δ en este tipo de estudios usualmente es pequeña y en cualquier caso más pequeña que la mínima diferencia aceptable desde el punto de vista clínico para una terapia considerada eficaz. El estudio de no inferioridad, con una sutil diferencia con respecto al anterior, busca mostrar que la nueva terapia no es peor que el tratamiento estándar o control activo. De este modo, el diseño de no inferioridad debe mostrar que la diferencia de efectos (nueva terapia – control activo) no es menor que un valor $-\Delta$ considerado el margen de no inferioridad (2).

El margen de no inferioridad

Para poder definir que la nueva terapia es no inferior es imprescindible fijar un margen para su efecto que lo delimite de la inferioridad. La elección de dicho margen, que no es sencilla ni tiene una respuesta

única, se fundamenta en el razonamiento clínico del problema y en algunas consideraciones estadísticas. Es necesario conocer con la mayor certeza posible, usualmente basada en una revisión sistemática de la literatura, el efecto que tuvo el tratamiento estándar comparado con el placebo en estudios previos, y según esta magnitud estimada del efecto original se elegirá la fracción *(f)* que se desee conservar en el nuevo tratamiento. Generalmente para esa valoración se tiene en cuenta la magnitud de no inferioridad que se está dispuesto a aceptar a cambio de otras ventajas de la terapia nueva como seguridad, tolerabilidad, costo o conveniencia. Un ejemplo es el de los estudios de tratamientos oncológicos cuando se evalúa mortalidad, para lo cual la Administración de Drogas y Alimentos de los Estados Unidos (FDA, por su sigla en inglés) sugiere una *f* de 0,5 (3). Con base en estos lineamientos, para la mayoría de estudios de no inferioridad se considera que dicho margen no debe ser menor del 50% del efecto que hubiese logrado la terapia estándar si pudiera compararse de manera simultánea con el placebo en ese mismo estudio.

Los intervalos de confianza y la interpretación de los resultados

El intervalo de confianza (IC) es la medida de la precisión de los resultados de una investigación, y en los ECA la amplitud de dicho intervalo determina la confianza que se tiene en la eficacia del tratamiento: mientras más estrecho sea el IC, lo que usualmente resulta de un tamaño de muestra más grande, más confiable es la estimación del efecto que se observa en el estudio. Usualmente se calcula un IC del 95% (IC 95%), cuya interpretación es que dentro de los límites estimados para dicho intervalo se encontrará la verdadera diferencia en promedio en 95 de 100 hipotéticos ECA idénticos al actual. Si el IC 95% del resultado de un ECA incluye el valor equivalente a ningún efecto (cero en diferencias absolutas como la reducción absoluta del riesgo, RAR, y uno para medidas relativas como el riesgo relativo, RR), entonces no se encontró una diferencia significativa entre los dos tratamientos. Pero de lo anterior no se puede concluir que los efectos de las terapias sean iguales. Puede existir una diferencia que el ECA fue incapaz de detectar a causa de un tamaño de muestra insuficiente y por tanto de una alta probabilidad de error de tipo II, lo que se denomina un bajo poder del estudio. En los estudios de superioridad también debe considerarse una diferencia clínica mínimamente aceptable o margen de superioridad, por encima del cual el tratamiento nuevo se considera realmente superior con respecto a la comparación. De conformidad con lo anterior, el límite inferior de eficacia potencial de acuerdo con el IC 95% debe sobrepasar dicho margen de superioridad (tratamiento nuevo – tratamiento actual = + Δ).

En los estudios de no inferioridad o equivalencia, el límite del IC 95% del efecto del nuevo tratamiento no debe cruzar el margen de no inferioridad (tratamiento nuevo – control activo = - Δ). Aunque existe la posibilidad de que el IC 95% cruce el margen de superioridad (tratamiento nuevo – control activo = + Δ), la posibilidad de interpretar este resultado como que el nuevo tratamiento es realmente superior depende de las consideraciones que se hicieron al calcular el tamaño de muestra y de la magnitud de dicha diferencia, ya que el diseño de estos estudios usualmente no permite sacar ese tipo de conclusiones. En la figura 1 se muestran las diferencias potencialmente observables (efecto del tratamiento nuevo – efecto del control activo: el cero representa dos tratamientos iguales, los valores negativos representan una mayor eficacia del control activo y los valores positivos representan una mayor eficacia del tratamiento nuevo), con su respectivo intervalo de confianza de 95%, en diversos resultados de estudios de no inferioridad y equivalencia: (A) La nueva terapia es significativamente mejor que el control activo, debido a que los límites del IC 95% siempre muestran valores positivos y el límite de superioridad había sido considerado previamente en el cálculo del tamaño de muestra. (B) La terapia nueva se considera equivalente al control activo, pues los límites del IC 95% están dentro del intervalo establecido (- Δ + Δ). En el caso de haberse establecido únicamente un margen de no inferioridad (- Δ), la terapia nueva se puede considerar no inferior al tratamiento estándar, pues el límite inferior del IC 95% para su eficacia está por encima de dicho margen de no inferioridad. (C) No se puede concluir que la nueva terapia sea no inferior o equivalente al control activo, ya que los límites de estimación de su efecto atraviesan también ambos límites del intervalo establecido. (D) La nueva terapia es inferior al control activo porque ambos límites del IC 95% muestran valores negativos, pero

la magnitud de dicha diferencia podría no ser clínicamente importante. Tampoco es posible establecer la no inferioridad porque el límite inferior del IC 95% se superpone con el margen establecido. (E) La terapia nueva es significativamente inferior a la terapia estándar, dado que ambos límites del IC 95% son menores de lo que se consideró margen de no inferioridad.

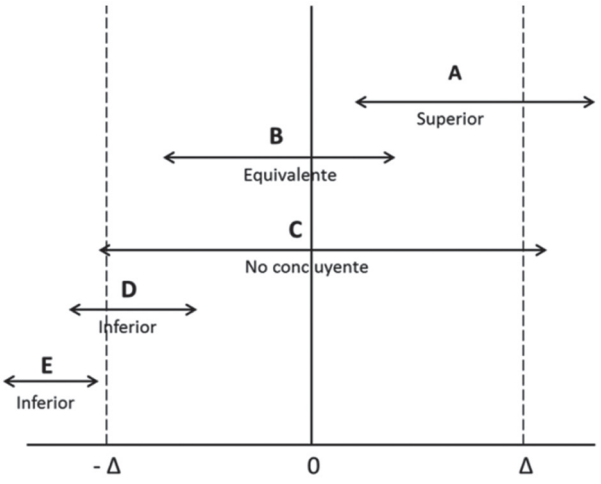

Figura 1. Resultados potenciales en estudios de no inferioridad/equivalencia

El principio del análisis por intención de tratar

Una consideración adicional acerca del análisis y la interpretación de los resultados es la concerniente al principio del análisis por intención de tratar (ITT). En los ECA es fundamental que el análisis de los resultados "proteja" las ventajas de la asignación aleatoria en cuanto a la comparabilidad de los grupos de tratamiento para evitar que los cambios en la adherencia, la exclusión de pacientes o cualquier suceso posterior a la asignación puedan sesgar los resultados. El ITT, de acuerdo con lo anterior, presume que lo que ocurre con la no adherencia o los cambios en el tratamiento se presenta de manera similar en los grupos tratados o no tratados, una situación denominada en investigación epidemiológica "clasificación inadecuada que ocurre de manera no diferencial" *(non differential misclassification)*. Esta clasificación inadecuada, por ocurrir de igual manera en los grupos, tiende a diluir cualquier potencial asociación y a mostrar las estimaciones más cercanas al punto de no efecto, es decir, una diferencia absoluta de cero o un riesgo relativo de 1. Esta tendencia a "borrar" las diferencias en el efecto que se produce con el uso del ITT se considera en cierta forma una ventaja en los estudios de superioridad, ya que si esta estimación conservadora de la eficacia de una intervención muestra diferencias a favor de la misma, es razonable suponer que la eficacia real del tratamiento pudiese ser aún mayor. En los estudios de no inferioridad o equivalencia, por el contrario, este mismo principio de ITT podría actuar de manera paradójica, ya que al mostrar estimaciones más cercanas al punto de no efecto está logrando "potenciar" la no inferioridad o la equivalencia del nuevo tratamiento con respecto al estándar, lo que puede llegar en casos extremos a sesgar los resultados del estudio. En los estudios de no inferioridad o equivalencia, por tanto, se deben presentar los resultados con base en el ITT, pero también complementados con el análisis de los pacientes que siguieron estrictamente el protocolo del estudio, lo que se ha llamado análisis por protocolo (PP). En algunos casos podría ser necesario mostrar también el análisis de acuerdo con los cambios de grupo o de tratamiento de los participantes, lo que se ha llamado análisis por tratamiento recibido (AT, por la sigla en inglés de *as treated*).

Un ejemplo reciente de un estudio de no inferioridad que resume las anteriores características es el "Co-formulated elvitegravir, cobicistat, emtricitabine, and tenofovir disoproxil fumarate versus ritonavir-boosted atazanavir plus co-formulated emtricitabine and tenofovir disoproxil fumarate for initial treatment of HIV-1 infection: a randomised, double-blind, phase 3, non-inferiority trial", llevado a cabo por DeJesus y colaboradores con el objetivo de comparar la seguridad y eficacia de los dos regímenes mencionados (EVG/COBI/FTC/TDF versus ATV/RTV+FTC/TDF) como tratamiento inicial de la infección por VIH-1(4). Se estableció la necesidad de una población de estudio de 700 pacientes como el tamaño de muestra con un poder de 95% para establecer no inferioridad; asumiendo una tasa de respuesta global, es decir, de supresión viral completa a la semana 48 de 79,5%, y un margen de no inferioridad de 12% en dicha proporción de supresión viral. Los análisis por intención de tratar (ITT) y por protocolo (PP) mostraron los resultados que se presentan en la tabla 1. En ambos casos se observa que los límites del IC 95% respetan el margen de no inferioridad del 12% que se había definido.

Tabla 1. Proporción de supresión viral completa a la semana 48, por grupos de tratamiento y tipo de análisis, en pacientes con infección por VIH-1

Tipo de análisis	EVG/COBI/FTC/TDF	ATV/RTV+FTC/TDF	Diferencia ajustada[a] (IC 95%)
ITT	316/353 (89,5%)	308/355 (86,8%)	3% (–1,9; 7,8)
PP	310/318 (97,5%)	303/310 (97,7%)	-0,1% (–2,6; 2,4)

[a]Ajustada de acuerdo con los estratos de carga viral inicial: mayor o menor de 100.000 copias por mL

Análisis crítico de un artículo de no inferioridad

A continuación se presenta un resumen de la guía de lectura crítica propuesta para este tipo de investigaciones (5).

¿Son los resultados válidos?

- ¿Los grupos de la nueva terapia y el tratamiento estándar iniciaron con el mismo pronóstico?

Como en los estudios clásicos de superioridad, los grupos que se comparan en el ECA de no inferioridad o equivalencia deben ser similares en términos de los factores relacionados con el pronóstico. Es decir, que cualquier diferencia o no diferencia en los resultados debería ser explicada únicamente por los efectos de las intervenciones en estudio. Para lo anterior, no solamente deben ser explícitos el procedimiento que se usó para la asignación aleatoria y el mecanismo para ocultar dicha asignación, sino que también debe ser verificable en los resultados del estudio el comportamiento en la línea de base de los principales determinantes del pronóstico, de acuerdo con los grupos de tratamiento que se comparan.

- ¿Se mantuvieron equilibrados los factores pronósticos durante el desarrollo de la investigación?

El complemento de la comparabilidad inicial evaluada en el anterior punto es que dicho balance se mantenga a través de todo el desarrollo del ECA y el seguimiento de los pacientes. Para esto, es necesario mantener el cegamiento o enmascaramiento de la mayor cantidad posible de participantes en la investigación: los pacientes, los clínicos, los encargados del seguimiento y recolección de datos, los evaluadores de los desenlaces y los analistas de los datos.

- ¿Se mantuvieron equilibrados los factores pronósticos al final de la investigación?

Para garantizar que el balance que pretende la asignación aleatoria se mantenga hasta el final, es necesario tener a los mismos pacientes que iniciaron el estudio también al final del mismo para su análisis, y específicamente en los mismos grupos a los que originalmente fueron asignados. Es decir, que se debe verificar que se presentaron mínimas o ninguna pérdida de seguimiento, y que el análisis principal de los resultados se hace con el ITT.

- ¿Se protegieron los investigadores de una conclusión de no inferioridad injustificada?

La característica más importante en este punto, que es específico para los ECA de no inferioridad o equivalencia, es que el tratamiento estándar o control activo empleado para comparación tenga realmente la eficacia esperada en el tratamiento de la enfermedad. Un control activo que sea subóptimo puede mostrar una aparente no inferioridad del nuevo tratamiento en los resultados del estudio. Esta disminución del efecto del tratamiento estándar puede ocurrir por reclutar pacientes con menos adherencia o menor respuesta a la terapia, con menor gravedad de su enfermedad, por suministrar el tratamiento de manera inadecuada o incompleta, o por terminar el seguimiento antes de observar la respuesta completa a la intervención. Una manera de explorar si el efecto del tratamiento estándar se ha preservado en el ECA de no inferioridad es comparar la frecuencia de los desenlaces en dicho estudio con lo que se ha visto en los ECA previos que han evaluado el mismo tratamiento.

- ¿Los investigadores analizaron a los pacientes según el tratamiento que recibieron, al igual que según los grupos a los cuales fueron asignados?

El ITT busca preservar las ventajas de la asignación aleatoria, pero en los estudios de no inferioridad dicho análisis puede subestimar el efecto del tratamiento estándar y de este modo llevar a una inferencia equivocada acerca de la no inferioridad del tratamiento nuevo. Si los resultados del análisis por protocolo (PP) son consistentes con los del ITT, y están por encima del margen de no inferioridad, se puede dar más validez a los resultados del estudio.

¿Cuáles fueron los resultados?

Los resultados en este tipo de estudios deben concentrarse en los siguientes puntos: 1) la diferencia entre el tratamiento nuevo y el estándar en cuanto al desenlace primario de eficacia que es la meta final del tratamiento para esa enfermedad; 2) los efectos adversos o las desventajas que deberían favorecer al tratamiento nuevo sobre el estándar; y 3) la confirmación en los resultados de que el tratamiento estándar se administró de manera óptima a la población correcta.

¿Cómo se pueden aplicar estos resultados al cuidado de los pacientes?

- ¿Fueron los pacientes del estudio similares a mi paciente?

La definición de la población de estudio del ECA, es decir, sus criterios de inclusión y exclusión, los procedimientos y sitios de reclutamiento de los participantes, así como la descripción final que se haga de las principales características de dicha población, son los elementos básicos para definir el paciente en el cual podrían potencialmente aplicarse las conclusiones de la investigación.

- ¿Se consideraron todos los desenlaces de importancia en los pacientes?

Los resultados de un ECA son aplicables en la medida en que sean explícitos para los interesados, médicos y pacientes, todos los efectos o desenlaces potenciales derivados de la intervención que se está evaluando.

- ¿Son mayores las ventajas del nuevo tratamiento frente al potencial daño y a los costos?

Este criterio es fundamental para decidir acerca de la utilidad y aplicación de un estudio de no inferioridad, y está basado principalmente en el juicio clínico acerca de qué tan razonable fue el margen de no inferioridad utilizado para el diseño y el análisis del estudio. Un tratamiento nuevo podría mostrar en su eficacia (medida como RAR o como RR) un límite del intervalo de confianza que sugiere que dicho efecto podría ser, aunque superior al placebo, menor del 50% del efecto que se espera del control activo. En este caso, la aceptación de dicho tratamiento nuevo como alterativa al estándar estaría condicionada a tener evidentes ventajas en cuanto a efectos adversos, riesgos, comodidad para el paciente o costos.

REFERENCIAS BIBLIOGRÁFICAS

1. Christensen E. Methodology of superiority vs. equivalence trials and non-inferiority trials. J Hepatol. 2007 May;46(5):947–54.
2. Head SJ, Kaul S, Bogers AJJC, Kappetein AP. Non-inferiority study design: lessons to be learned from cardiovascular trials. Eur Hear. J. 2012 Jun;33(11):1318–24.
3. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. Ann Intern Med. 2006 Jul 4;145(1):62–9.
4. DeJesus E, Rockstroh JK, Henry K, Molina J-M, Gathe J, Ramanathan S, et al. Co-formulated elvitegravir, cobicistat, emtricitabine, and tenofovir disoproxil fumarate versus ritonavir-boosted atazanavir plus co-formulated emtricitabine and tenofovir disoproxil fumarate for initial treatment of HIV-1 infection: a randomised, double-. Lancet. 2012 Jun 30;379(9835):2429–38.
5. Mulla SM, Scott IA, Jackevicius CA, You JJ, Guyatt GH. How to use a noninferiority trial: users' guides to the medical literature. JAMA. 2012 Dec 26;308(24):2605–11.