ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Julio 2018

FORMACIÓN MÉDICA

Rey de reyes. El ensayo clínico aleatorizado

Molina Arias M

Hospital Infantil Universitario La Paz, Madrid.

Resumen

El ensayo clínico es un estudio de intervención, analítico, de dirección anterógrada y temporalidad concurrente, con muestreo de cohorte cerrada con control de la exposición. Es el diseño más adecuado para valorar la eficacia de un tratamiento y el que proporciona una evidencia de mayor calidad para demostrar la relación de causalidad entre la intervención y los resultados observados. Repasamos su estructura, sus medidas de asociación e impacto y algunas variaciones de su diseño básico.

Introducción

El ensayo clínico es un estudio de intervención, analítico, de dirección anterógrada y temporalidad concurrente, con muestreo de cohorte cerrada con control de la exposición. Es el diseño más adecuado para valorar la eficacia de un tratamiento y el que proporciona una evidencia de mayor calidad para demostrar la relación de causalidad entre la intervención y los resultados observados. Repasamos su estructura, sus medidas de asociación e impacto y algunas variaciones de su diseño básico.

No cabe duda de que a la hora de realizar un trabajo de investigación en biomedicina podemos elegir entre un gran número de diseños posibles, todos ellos con sus ventajas e inconvenientes. Pero en esta corte tan diversa y poblada, entre malabaristas, sabios, jardineros y flautistas púrpuras, reina por encima de todos el verdadero Rey Carmesí de la epidemiología: **el ensayo clínico aleatorizado**.

El ensayo clínico es un estudio de intervención, analítico, de dirección anterógrada y temporalidad concurrente, con muestreo de cohorte cerrada con control de la exposición. En un ensayo se selecciona una muestra de una población y se divide al azar en dos grupos. Uno de los grupos (grupo de intervención) sufre la intervención que queremos estudiar, mientras que el otro (grupo de control) nos sirve de referencia para comparar los resultados. Tras un periodo de seguimiento determinado se analizan los resultados y se comparan las diferencias entre los dos grupos. Podemos así evaluar los beneficios de los tratamientos o intervenciones al mismo tiempo que controlamos los sesgos de otros tipos de estudios: la aleatorización favorece que los posibles factores de confusión, conocidos o no, se repartan de forma uniforme entre los dos grupos, de forma que si al final detectamos alguna diferencia, esta tiene que ser debida a la intervención en estudio. Esto es lo que nos permite establecer una relación causal entre exposición y efecto.

Por lo dicho hasta ahora, se comprende fácilmente que el ensayo clínico aleatorizado sea el diseño **más adecuado** para valorar la eficacia de cualquier intervención en medicina y es

el que proporciona, como ya hemos comentado, una evidencia de mayor calidad para demostrar la relación de causalidad entre la intervención y los resultados observados.

Pero para disfrutar de todos estos beneficios es necesario ser escrupuloso en el planteamiento y metodología de los ensayos. Existen listas de verificación publicadas por sabios que entienden mucho de estos temas, como es el caso de la lista CONSORT, que nos pueden ayudar a valorar la calidad del diseño del ensayo. Pero entre todos estos aspectos, reflexionemos un poco sobre aquellos que son cruciales para la validez del ensayo clínico.

Tal como veis en la figura 1, todo empieza con una laguna de conocimiento que nos lleva a formular una pregunta clínica estructurada. El único objetivo del ensayo debe ser responder a esta pregunta y basta con que se responda de forma adecuada a una sola pregunta. Desconfiad de los ensayos clínicos que tratan de responder a muchas preguntas, ya que, en muchas ocasiones, al final no responden bien a ninguna. Además, el planteamiento debe basarse en lo que los inventores de jerga metodológica llaman el principio de incertidumbre (la *equipoise* de lo que hablan inglés), que no quiere decir más que, en el fondo de nuestro corazón, desconocemos de verdad cuál de las dos intervenciones es más beneficiosa para el paciente (habría que ser un poco perro desde el punto de vista ético para realizar una comparación si ya sabemos con seguridad cuál de las dos intervenciones es mejor).

Figura 1. ECA paralelo

A continuación debemos elegir cuidadosamente la muestra sobre la que realizaremos el ensayo. Idealmente, todos los miembros de la población deberían tener la misma probabilidad no solo de ser elegidos, sino también de acabar en cualquiera de las dos ramas del ensayo. Aquí nos encontramos con un pequeño dilema. Si somos muy estrictos con los criterios de inclusión y exclusión la muestra será muy homogénea y la validez interna del estudio saldrá fortalecida, pero será más difícil extender los resultados a la población general (esta es la actitud explicativa de selección de la muestra). Por otra parte, si no somos tan rígidos los resultados se parecerán más a los de la población general, pero puede verse comprometida la validez interna del estudio (esta es la actitud pragmática).

La **aleatorización** (¿quién ha dicho randomización?) **es uno de los puntos clave** del ensayo clínico. Es la que nos asegura que podemos comparar los dos grupos, ya que tiende a distribuir por igual las variables conocidas y, más importante, también las desconocidas entre los dos grupos. Pero no nos relajemos demasiado: este reparto no está en absoluto garantizado, solo es más probable que ocurra si aleatorizamos de forma correcta, así que siempre deberemos comprobar la homogeneidad de los dos grupos, sobre todo con muestras pequeñas.

Además, la aleatorización nos permite realizar de forma adecuada el enmascaramiento, con lo que realizamos una medición no sesgada de la variable de respuesta, evitando los sesgos de información. Estos resultados del grupo de intervención los podemos comparar con los del grupo control de tres formas. Una de ellas es comparar con un placebo. El placebo debe ser un preparado de características físicas indistinguibles del fármaco de intervención pero sin sus efectos

farmacológicos. Esto sirve para controlar el efecto placebo (que depende de la personalidad del paciente, de sus sentimientos hacia a la intervención, de su cariño por el equipo investigador, etc), pero también los efectos secundarios que son debidos a la intervención y no al efecto farmacológico (pensemos, por ejemplo, en el porcentaje de infecciones locales en un ensayo con medicación administrada por vía intramuscular).

La otra forma de comparar es con el tratamiento aceptado como más eficaz hasta el momento. Si existe un tratamiento que funciona, lo lógico (y más ético) es que lo usemos para investigar si el nuevo aporta beneficios. También suele ser el método de comparación habitual en los estudios de equivalencia o de no-inferioridad. Por último, la tercera posibilidad es comparar con la no intervención, aunque en realidad esto es una forma rebuscada de decir que solo se le aplican los cuidados habituales que recibiría cualquier paciente en su situación clínica.

Es imprescindible que todos los participantes en el ensayo sean sometidos a la misma pauta de seguimiento, que debe ser lo suficientemente prolongado como para permitir que se produzca la respuesta esperada. Deben detallarse y analizarse todas las pérdidas que se produzcan durante el seguimiento, ya que pueden comprometer la validez y la potencia del estudio para detectar diferencias significativas. ¿Y qué hacemos con los que se pierden o acaban en una rama diferente a la asignada? Si son muchos, lo más razonable puede ser rechazar el estudio. Otra posibilidad es excluirlos y hacer como si no hubiesen existido nunca, pero podemos sesgar los resultados del ensayo. Una tercera posibilidad es incluirlos en el análisis en la rama del ensayo en la que han participado (siempre hay alguno que se confunde y se toma lo que no le toca), lo que se conoce como análisis por tratamiento o análisis por protocolo. Y la cuarta, y última opción que tenemos, es analizarlos en la rama que se les asignó inicialmente con independencia de lo que hayan hecho durante el estudio. Esto se denomina análisis por intención de tratar, y es la única de las cuatro posibilidades que nos permite conservar todos los beneficios que previamente nos había proporcionado la aleatorización.

Como **fase final**, nos quedaría el análisis y comparación de los datos para extraer las conclusiones del ensayo, utilizando para ello las medidas de asociación y medidas de impacto oportunas que, en el caso del ensayo clínico, suelen ser la tasa de respuesta, el riesgo relativo (RR), la reducción relativa del riesgo (RRR), la reducción absoluta del riesgo (RAR) y el número necesario a tratar (NNT). Vamos a verlos con un ejemplo.

Imaginemos que realizamos un ensayo clínico en el que probamos un antibiótico nuevo (llamémosle A para no calentarnos mucho la cabeza) para el tratamiento de una infección grave de la localización que nos interese estudiar. Aleatorizamos los pacientes seleccionados y les damos el fármaco nuevo o el tratamiento habitual (nuestro grupo de control), según les corresponda por azar. Al final, medimos en cuántos de nuestros pacientes fracasa el tratamiento (el evento que queremos evitar).
De los 100 pacientes que reciben el fármaco A, 36 presentan el evento a evitar. Por tanto, podemos concluir que el riesgo o incidencia del evento en los expuestos (Ie) es de 0,36 (36 de cada 100, en tanto por uno). Por otra parte, 60 de los 100 controles (los llamamos el grupo de no expuestos) han presentado el suceso, por lo que rápidamente

calculamos que el riesgo o incidencia en los no expuestos (Io) es de 0,6.

A simple vista ya vemos que el riesgo es distinto en cada grupo, pero como en la ciencia hay que medirlo todo, podemos dividir los riesgos entre expuestos y no expuestos, obteniendo así el denominado riesgo relativo (RR = Ie/Io). Un RR = 1 significa que el riesgo es igual en los dos grupos. Si el RR > 1 el evento será más probable en el grupo de expuestos (la exposición que estemos estudiando será un factor de riesgo para la producción del evento) y si RR está entre 0 y 1, el riesgo será menor en los expuestos. En nuestro caso, RR = 0,36/0,6 = 0,6. Es más sencillo interpretar los RR > 1. Por ejemplo, un RR de 2 quiere decir que la probabilidad del evento es dos veces mayor en el grupo expuesto. Siguiendo el mismo razonamiento, un RR de 0,3 nos diría que el evento es una tercera parte menos frecuente en los expuestos que en los controles. Podéis ver en la figura 2 cómo se calculan estas medidas.

	Evento	No evento
Fármaco A	36	64
Control	60	40

Ie = 0,36
Io = 0,6
RR = 0,36/0,6 = 0,6
RRR = (0,36 – 0,6) / 0,6 = 0,4
RAR = 0,36 – 0,6 = 0,24
NNT = 1 / 0,24 = 4

	Evento	No evento
Fármaco B	3	97
Control	5	95

Ie = 0,03
Io = 0,05
RR = 0,03/0,05 = 0,6
RRR = (0,03 – 0,05) / 0,05 = 0,4
RAR = 0,03 – 0,05 = 0,02
NNT = 1 / 0,02 = 50

Figura 2. Medidas ensayo clínico

Pero lo que a nosotros nos interesa es saber cuánto disminuye el riesgo del evento con nuestra intervención para estimar cuánto esfuerzo hace falta para prevenir cada uno. Para ello podemos calcular la **RRR y la RAR**. La RRR es la diferencia de riesgo entre los dos grupos respecto del control (RRR = [Ie-Io]/Io). En nuestro caso es de 0,4, lo que quiere decir que la intervención probada disminuye el riesgo un 60% respecto al tratamiento habitual.

La RAR es más sencilla: es la resta entre los riesgos de expuestos y controles (RAR = Ie – Io). En nuestro caso es de 0,24 (prescindimos del signo negativo), lo que quiere decir que de cada 100 pacientes que tratemos con el nuevo fármaco se producirán 24 eventos menos que si hubiésemos utilizado el tratamiento control. Pero aún hay más: podemos saber cuántos tenemos que tratar con el fármaco nuevo para evitar un evento con solo hacer la regla de tres (24 es a 100 como 1 es a x) o, más fácil de recordar, calculando el inverso de la RAR. Así, el NNT = 1/RAR = 4,1. En nuestro caso tendríamos que tratar a cuatro pacientes para evitar un suceso adverso. El contexto nos dirá siempre la importancia clínica de esta cifra.

Como veis, la RRR, aunque es técnicamente correcta, tiende a magnificar el efecto y no nos cuantifica claramente el esfuerzo a realizar para obtener los resultados. Además, puede ser similar en situaciones diferentes con implicaciones clínicas totalmente distintas. Veámoslo con otro ejemplo que también os muestro en la figura 2. Supongamos otro ensayo con un fármaco B en el que obtenemos tres eventos en los 100 tratados y cinco en los 100 controles. Si hacéis los cálculos, el RR es de 0,6 y la RRR de 0,4, igual que en el ejemplo anterior, pero si calculáis la RAR veréis que es muy diferente (RAR = 0,02), con un NNT de 50. Se ve claramente que el esfuerzo para evitar un evento es mucho mayor (cuatro frente a 50) a pesar de que coincidan el RR y la RRR.

Así que, llegados a este punto, permitidme un consejo. Dado que con los datos necesarios para calcular la RRR es incluso más sencillo calcular la RAR (y el NNT), si en un trabajo científico os lo ocultan y solo os ofrecen

la RRR, desconfiad como del cuñado que os pone un queso curado para meteros el vino barato y preguntadle por qué no os pone mejor un pincho de jamón ibérico. Bueno, en realidad quería decir que os preguntéis por qué no os dan la RAR y la calculéis vosotros con los datos del trabajo.

Hasta ahora todo lo que hemos dicho hace referencia al diseño clásico de ensayo clínico en paralelo, pero el rey de los diseños tiene muchas caras y, con mucha frecuencia, podemos encontrar trabajos en los que se nos muestra de forma un poco diferente, lo que puede implicar que el análisis de los resultados tenga peculiaridades especiales.

Vamos a empezar con una de las variaciones más frecuentes. Si lo pensamos un momento, el diseño ideal sería aquel que nos permitiese experimentar en el mismo individuo el efecto de la intervención de estudio y de la de control (el placebo o el tratamiento estándar), ya que el ensayo en paralelo es una aproximación que supone que los dos grupos responden igual a las dos intervenciones, lo que siempre supone un riesgo de sesgo que tratamos de minimizar con la aleatorización. Si tuviésemos una máquina del tiempo podríamos probar la intervención en todos, anotar lo que pasa, dar marcha atrás en el tiempo y volver a repetir el experimento con la intervención de control. Así podríamos comparar los dos efectos. El problema, los más atentos ya lo habréis imaginado, es que la máquina del tiempo no se ha inventado todavía.

Pero lo que sí se ha inventado es el ensayo clínico cruzado (el cross-over, para los que sepan inglés), en el que cada sujeto es su propio control. Como podéis ver en la figura 3, en este tipo de ensayo cada sujeto es aleatorizado a un grupo, se le somete a la intervención, se deja pasar un periodo de lavado o blanqueo y se le somete a la otra intervención. Aunque esta solución no es tan elegante como la de la máquina del tiempo, los defensores de los ensayos cruzados se basan en que la variabilidad dentro de cada individuo es menor que la interindividual, con lo cual la estimación puede ser más precisa que la del ensayo en paralelo y, en general, se necesitan tamaños muestrales menores. Eso sí, antes de utilizar este diseño hay que hacer una serie de consideraciones. Lógicamente, el efecto de la primera intervención no debe producir cambios irreversibles ni ser muy prolongado, porque afectaría el efecto de la segunda. Además, el periodo de lavado tiene que ser lo suficientemente largo para evitar que quede ningún efecto residual de la primera intervención.

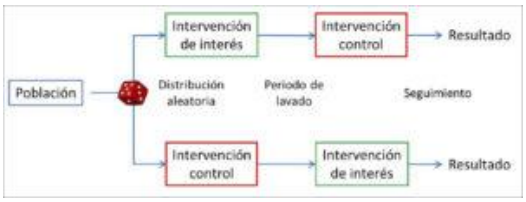

Figura 3. Ensayo cruzado

También hay que considerar si el orden de las intervenciones puede afectar el resultado final (efecto secuencia), con lo que solo serían válidos los resultados de la primera intervención. Otro problema es que, al tener mayor duración, las características del paciente pueden cambiar a lo largo del estudio y ser diferentes en los dos periodos (efecto periodo). Y, por último, ojo con las pérdidas durante el estudio, más frecuentes en estudios más largos y que tienen en los ensayos cruzados mayor repercusión sobre los resultados finales que en los ensayos en paralelo.

Imaginemos ahora que queremos probar dos intervenciones (A y B) en la misma población. ¿Podemos hacerlo con un mismo ensayo y ahorrar costes de todo tipo? Pues sí, sí que podemos, solo tenemos que diseñar un ensayo clínico

factorial. En este tipo de ensayo, cada participante es sometido a dos aleatorizaciones consecutivas: primero se le asigna a la intervención A o al placebo (P) y, segundo, a la intervención B o al placebo, con lo que tendremos cuatro grupos de estudio: AB, AP, BP y PP. Como es lógico, las dos intervenciones deben actuar por mecanismos independientes para poder valorar los resultados de los dos efectos de forma independiente.

Habitualmente se estudian una intervención relacionada con una hipótesis más plausible y madura y otra con una hipótesis menos contrastada, asegurando que la evaluación de la segunda no influye sobre los criterios de inclusión y exclusión de la primera. Además, no es conveniente que ninguna de las dos opciones tenga muchos efectos molestos o sea mal tolerada, porque la falta de cumplimiento de un tratamiento suele condicionar el mal cumplimiento del otro. En casos en que las dos intervenciones no se muestren independientes, podrían estudiarse los efectos por separado (AP frente a PP y BP frente a PP), pero se pierden las ventajas del diseño y aumenta el tamaño de muestra necesario.

En otras ocasiones puede ocurrir que tengamos prisa por acabar el estudio cuanto antes. Imaginemos una enfermedad muy mala que mata la gente a montones y nosotros estamos probando un nuevo tratamiento. Querremos tenerlo disponible cuanto antes (si funciona, claro), así que cada cierto número de participantes nos pararemos y analizaremos y, en el caso de que podamos demostrar ya la utilidad del tratamiento, daremos el estudio por concluido. Este es el diseño que caracteriza al ensayo clínico secuencial. Recordad que en el ensayo en paralelo lo correcto es calcular previamente el tamaño de la muestra. En este diseño, de mentalidad más bayesiana, se establece un estadístico cuyo valor condiciona una regla de finalización explícita, con lo que el tamaño de la muestra depende de las observaciones previas. Cuando el estadístico alcanza el valor prefijado nos vemos con la suficiente confianza como para rechazar la hipótesis nula y finalizamos el estudio. El problema es que cada parón y análisis aumenta el error de rechazarla siendo cierta (error de tipo 1), por lo que no se recomienda hacer muchos análisis intermedios. Además, el análisis final de los resultados es complejo porque los métodos habituales no sirven, sino que hay utilizar otros que tengan en cuenta los análisis intermedios. Este tipo de ensayos es muy útil con intervenciones de efecto muy rápido, por lo que es frecuente verlos en estudios de titulación de dosis de opiáceos, hipnóticos y venenos semejantes.

Hay otras ocasiones en las que la aleatorización individual no tiene sentido. Pensemos que hemos enseñado a los médicos de un centro de salud una nueva técnica para informar mejor a sus pacientes y queremos compararla con la antigua. No podemos decir al mismo médico que informe a unos pacientes de una forma y a otros de otra, ya que habría muchas posibilidades de que las dos intervenciones se contaminaran una a otra. Sería más lógico enseñar a los médicos de un grupo de centros y no enseñar a los de otro grupo y comparar los resultados. Aquí lo que aleatorizaríamos son los centros de salud para formar o no a sus médicos. Este es el diseño de ensayo con asignación por grupos. El problema de este diseño es que no tenemos muchas garantías de que los participantes de los diferentes grupos se comporten de forma independiente, por lo que el tamaño de la muestra necesaria puede aumentar mucho si existe gran variabilidad entre los grupos y poca dentro de cada grupo. Además, hay que

hacer un análisis agregado de los resultados, ya que si se hace individual los intervalos de confianza se estrechan de forma artefactada y podemos encontrar significaciones estadísticas falsas. Lo habitual es calcular un estadístico sintético ponderado para cada grupo y hacer las comparaciones finales con él.

El último de la serie que vamos a tratar es el ensayo comunitario, en el cual la intervención se aplica a grupos de población. Al realizarse en condiciones reales sobre poblaciones tienen gran validez externa y permiten muchas veces recomendar medidas coste-eficientes basadas en sus resultados. El problema es que muchas veces es complicado establecer grupos de control, puede ser más difícil determinar el tamaño muestral necesario y es más complejo realizar inferencia causal a partir de sus resultados. Es el diseño típico para evaluar medidas de salud pública como la fluoración del agua, las vacunaciones, etc.

Acabo ya. La verdad es que esta entrada me ha quedado un poco larga (y espero que no demasiado insufrible), pero es que el Rey se lo merece. De todas formas, si pensáis que está todo dicho sobre ensayos clínicos no tenéis ni idea de todo lo que queda por decir sobre tipos de muestreos, de aleatorización, etc, etc, etc. Pero esa es otra historia…

Bibliografía

– Molina M, Ochoa C. Ensayo clínico (I). Definición. Tipos. Estudios cuasi experimentales. Evid Pediatr.2014; 10:52. (HTML)

– Molina M, Ochoa C. Ensayo clínico (V). Estrategias de análisis. Pérdidas de información. Análisis por intención de tratar. Evid Pediatr.2015;11:52. (HTML)

– Bakhai A, Chhabra A, Wang D. Endpoints. En: Wang D, Bakhai A (eds.). Clinical trials. A practical guide to design, analysis, and reporting. Chicago: Remedica; 2006. p. 37-46.

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología
Hospital Infantil Universitario La Paz, Madrid.

Aceptado para blog en febrero de 2018.