

El rey a examen

Todos sabemos que el ensayo clínico aleatorizado es el rey de los diseños metodológicos de intervención. Es el tipo de estudio epidemiológico que permite un mejor control de los errores sistemáticos o sesgos, ya que el investigador controla las variables del estudio y los participantes son asignados al azar entre las intervenciones que se comparan.

De esta forma, si dos grupos homogéneos que difieren solo en la intervención presentan alguna diferencia de interés durante el seguimiento, podremos afirmar con cierta confianza que esta diferencia se debe a la intervención, lo único que no tienen en común los dos grupos. Por este motivo, el ensayo clínico es el diseño preferente para responder a preguntas clínicas sobre intervención o tratamiento, aunque siempre tendremos que ser prudentes con la evidencia generada por un solo ensayo clínico, por muy bien realizado que esté. Cuando realizamos una revisión sistemática sobre ensayos clínicos aleatorizados sobre la misma intervención y los combinamos en un metanálisis, las respuestas que obtenemos tendrán más fiabilidad que las obtenidas de un solo estudio. Por eso hay quien afirma que el diseño ideal para responder preguntas de tratamiento no es el ensayo clínico, sino el metanálisis de ensayos clínicos.

En cualquier caso, como las revisiones sistemáticas valoran sus estudios primarios de forma individual y como lo más habitual es que encontremos ensayos individuales y no revisiones sistemáticas, conviene saber hacer una buena lectura crítica para poder sacar conclusiones. En efecto, no podemos relajarnos cuando veamos que un artículo nos cuenta un ensayo clínico y darlo por bueno. El ensayo clínico puede también contener sus trampas y argucias, por lo que, como con cualquier otro tipo de trabajo, será buena práctica realizar la lectura crítica del mismo, basándonos en nuestros [tres pilares](#) habituales: validez, importancia y aplicabilidad.

Como siempre, a la hora de estudiar el rigor científico o [VALIDEZ](#) (validez interna), nos fijaremos primero en una serie de criterios primarios imprescindibles. Si estos no se cumplen, mejor no perder el tiempo con el trabajo y buscar otro más provechoso.

¿Existe una pregunta clínica claramente definida? En su origen, el ensayo debe estar diseñado para responder a una pregunta clínica estructurada sobre tratamiento, motivada por una de nuestras múltiples lagunas de conocimiento. Se debe plantear una hipótesis de trabajo con sus correspondientes hipótesis nula y alternativa, a ser posible sobre un tema relevante desde el punto de vista clínico. Es preferible que el estudio trate de responder solo a una pregunta. Cuando se quiere responder a varias

suele complicarse el estudio en exceso para acabar no contestando ninguna de forma completa y adecuada.

¿Se realizó la asignación de forma aleatoria? Como ya hemos dicho, para poder afirmar que las diferencias entre los grupos se deben a la intervención es necesario que sean homogéneos. Esto se consigue asignando los pacientes al azar, única forma de controlar las variables confusoras conocidas y, más importante, también las que desconocemos. Si los grupos fueran distintos y atribuyésemos la diferencia únicamente a la intervención podríamos incurrir en un sesgo de confusión. El ensayo debe contener la habitual e imprescindible tabla 1 con la frecuencia de aparición de las variables demográficas y de confusión de ambas muestras para estar seguros de que los grupos son homogéneos. Un error frecuente es buscar las diferencias entre los dos grupos y valorarlas según su p , cuando sabemos que la p no mide homogeneidad. Si los hemos repartido al azar, cualquier diferencia que observemos se deberá obligatoriamente al azar (no necesitaremos una p para saberlo). El tamaño muestral no está pensado para discriminar entre las variables demográficas, por lo que una p no significativa puede indicar simplemente que la muestra es pequeña para verla. Por otro lado, cualquier mínima diferencia puede alcanzar significación estadística si la muestra es lo suficientemente grande. Así que olvidaos de la p : si hay alguna diferencia, lo que hay que hacer es valorar si tiene la relevancia clínica suficiente como para poder haber influido en los resultados o, dicho de forma más elegante, tendremos que controlar las covariables no equilibradas durante la aleatorización. Afortunadamente, cada vez es más raro encontrar las tablas de los grupos de estudio con las p de comparación entre los grupos de intervención y control.

Pero no basta con que el estudio esté aleatorizado (aprovecho para decir que randomizado es un barbarismo que no existe en lengua castellana), sino que hay que considerar también si la secuencia de aleatorización se hizo de forma correcta. El método utilizado debe garantizar que todos los componentes de la población seleccionada tengan la misma probabilidad de ser elegidos, por lo que se prefieren las tablas de números aleatorios o secuencias generadas por ordenador. La aleatorización debe ser oculta, de forma que no se pueda saber a qué grupo va a pertenecer el siguiente participante. Por eso gustan tanto los sistemas centralizados vía telefónica o a través de Internet. Y aquí pasa algo muy curioso: resulta que es bien conocido que la aleatorización produce muestras de diferente tamaño, sobre todo si las muestras son pequeñas, motivo por el que a veces se usan muestras aleatorizadas por bloques balanceados en tamaño. Y yo os pregunto, ¿cuántos estudios habéis leído con el mismo número de participantes en las dos ramas y que afirmaban ser aleatorizados? Desconfiad si veis grupos iguales, sobre todo si son pequeños, y no os dejéis engañar: siempre podéis utilizar una de las múltiples calculadoras

de probabilidad binomial disponibles en la Red para saber cuál es la probabilidad de que el azar genere los grupos que los autores os presentan (hablamos siempre de aleatorización simple, no por bloques, conglomerados, minimización u otras técnicas). Os sorprenderá lo que encontraréis.

También es importante que el seguimiento haya sido lo suficientemente largo y completo, de forma que el estudio dure lo suficiente para poder observar la variable de resultado y que todo participante que entre en el estudio sea tenido en cuenta al finalizar. Como regla de andar por casa, si las pérdidas superan el 20%, se admite que la validez interna del estudio puede estar comprometida.

Siempre tendremos que analizar la naturaleza de las pérdidas durante el seguimiento, sobre todo si estas son elevadas. Hay que tratar de dilucidar si las pérdidas son aleatorias o si están en relación con alguna variable concreta (lo cual sería mal asunto) y estimar qué efecto han podido tener en los resultados del ensayo. Lo más habitual suele ser adoptar el llamado escenario del peor de los casos: se supone que todas las pérdidas del grupo control han ido bien y todas las del grupo de intervención han ido mal y se repite el análisis para comprobar si las conclusiones se modifican, en cuyo caso la validez del estudio quedaría seriamente comprometida. El último aspecto importante es considerar si los pacientes que no han recibido el tratamiento previamente asignado (siempre hay alguno que no se entera y mete la pata) se han analizado según la intención de tratamiento, ya que es la única forma de preservar todos los beneficios que se obtienen con la aleatorización. Todo aquello que ocurra después de la aleatorización (como un cambio del grupo de asignación) puede influir en la probabilidad de que el sujeto experimente el efecto que estamos estudiando, por lo que es importante respetar este análisis por intención de tratar y analizar cada uno en el grupo en el que se asignó inicialmente.

Una vez comprobados estos criterios primarios, nos fijaremos en tres criterios secundarios que influyen en la validez interna. Habrá que comprobar que los grupos fueran similares al inicio del estudio (ya hemos hablado de la tabla con los datos de los dos grupos), que se llevó a cabo el enmascaramiento de forma adecuada como forma de control de sesgos y que los dos grupos fueron manejados y controlados de forma similar a excepción, claro está, de la intervención en estudio. Sabemos que el enmascaramiento o cegamiento permite minimizar el riesgo de sesgos de información, por lo que lo habitual es que los investigadores y los participantes desconozcan a qué grupo está asignado cada uno, lo que se conoce como doble ciego. En ocasiones, dada la naturaleza de la intervención (pensemos un grupo que se opera y otro no) será imposible enmascarar a investigadores y participantes, pero siempre podremos darle los datos enmascarados a la persona que realice el análisis de los resultados (el llamado evaluador ciego), con lo que se trata de paliar este problema.

Para resumir este apartado de la validez del ensayo, podemos decir que tendremos que comprobar que exista una clara definición de la población de estudio, de la intervención y del resultado de interés, que la aleatorización se haya hecho de forma adecuada, que se han tratado de controlar los sesgos de información mediante el enmascaramiento, que ha habido un seguimiento adecuado con control de las pérdidas y que el análisis haya sido correcto (análisis por intención de tratamiento y control de covariables no equilibradas por la aleatorización).

Una herramienta muy sencilla que nos puede ayudar también a valorar la validez interna de un ensayo clínico es la escala de Jadad, también llamada sistema de puntuación de calidad de Oxford. Jadad, un médico colombiano, ideó un sistema de puntuación con 7 preguntas. En primer lugar, 5 preguntas cuya respuesta afirmativa suma 1 punto:

1. ¿El estudio se describe como aleatorizado?
 2. ¿Se describe el método utilizado para generar la secuencia de aleatorización y este es adecuado?
 3. ¿El estudio se describe como doble ciego?
 4. ¿Se describe el método de enmascaramiento y este es adecuado?
 5. ¿Hay una descripción de las pérdidas durante el seguimiento?
- Finalmente, dos preguntas cuya respuesta negativa resta 1 punto:

1. ¿Es adecuado el método utilizado para generar la secuencia de aleatorización?
2. ¿Es adecuado el método de enmascaramiento?

Como veis, la escala de Jadad valora los puntos clave que ya hemos mencionado: aleatorización, enmascaramiento y seguimiento. Se considera un estudio riguroso desde el punto de vista metodológico el que tiene una puntuación de 5 puntos. Si el estudio tiene 3 puntos o menos, mejor lo utilizamos para envolver el bocadillo.

Pasaremos a continuación a considerar cuáles son los resultados del estudio para calibrar su **IMPORTANCIA** clínica. Habrá que determinar las variables medidas para ver si el trabajo expresa de forma adecuada la magnitud y la precisión de los resultados. Es importante, una vez más, no conformarnos con que nos inunden con múltiples *p* llenas de ceros. Recordad que la *p* solo nos indica la probabilidad de que estemos dando como buenas diferencias que solo existen por azar (o, dicho con elegancia, de cometer un [error de tipo 1](#)), pero que significación estadística no tiene porqué ser sinónimo de [relevancia clínica](#).

En el caso de variables continuas como tiempo de supervivencia, peso, tensión arterial, etc, lo habitual será expresar la magnitud de los resultados como diferencia de medias o de medianas, dependiendo de cuál sea la [medida de centralización](#) más adecuada. Sin embargo, en casos de variables dicotómicas (vivo o muerto, sano o enfermo, etc) se utilizarán el

riesgo relativo, su reducción relativa y absoluta y el número necesario a tratar (NNT). De todas ellas, la que mejor expresa la eficiencia clínica es siempre el [NNT](#). Cualquier trabajo digno de nuestra atención debe proporcionar estos datos o, en su defecto, la información necesaria para que podamos calcularlos.

Pero para permitir conocer una estimación más real de los resultados en la población necesitamos saber la precisión del estudio, y nada más fácil que recurrir a los [intervalos de confianza](#). Estos intervalos, además de la precisión, nos informan también de la significación estadística. Será estadísticamente significativo si el intervalo del riesgo relativo no incluye el uno y el de la diferencia de medias, el cero. En el caso de que los autores no nos los proporcionen, podemos utilizar una calculadora para obtenerlos, como las disponibles en la web de [CASPe](#).

Una buena forma de ordenar el estudio de la importancia clínica de un ensayo es estructurarlo en las cuatro Cs: valoración Cuantitativa (medidas de efecto y su precisión), valoración Cualitativa (relevancia desde el punto de vista clínico), valoración Comparativa (ver si los resultados son congruentes con los de otros estudios previos) y valoración de Coste-beneficio (este punto enlazaría con el siguiente apartado de la lectura crítica que tiene que ver con la aplicabilidad de los resultados del ensayo).

Para finalizar la lectura crítica de un artículo de tratamiento valoraremos su [APLICABILIDAD](#) (también llamada validez externa), para lo cual nos tendremos que preguntar si los resultados pueden generalizarse a nuestros pacientes o, dicho de otro modo, si existe alguna diferencia entre nuestros pacientes y los del estudio que impida la generalización de los resultados. Hay que tener en cuenta en este sentido que cuánto más estrictos sean los criterios de inclusión de un estudio, más difícil será generalizar sus resultados, comprometiéndose así su validez externa.

Pero, además, debemos considerar si se han tenido en cuenta todos los resultados clínicamente importantes, incluyendo efectos secundarios e indeseables. La variable de resultado medida debe ser importante para el médico y para el paciente. No hay que olvidar que el hecho de que demostrar que la intervención sea eficaz no significa obligatoriamente que sea beneficiosa para nuestros pacientes. Habrá que valorar también los efectos nocivos o molestos y estudiar el balance beneficios-costes-riesgos, así como las dificultades que puedan existir para aplicar el tratamiento en nuestro medio, las preferencias del paciente, etc.

Como es fácil de comprender, un estudio puede tener una gran validez metodológica y sus resultados tener gran importancia desde el punto de vista clínico y no ser aplicable a nuestros pacientes, ya sea porque nuestros pacientes sean diferentes a los del estudio, porque no se adapte a

sus preferencias o porque sea irrealizable en nuestro medio. Sin embargo, lo contrario no suele ocurrir: si la validez es deficiente o los resultados carecen de importancia, difícilmente nos plantearemos aplicar las conclusiones del estudio a nuestros pacientes.

Para terminar, recomendaros que utilizéis alguna de las herramientas disponibles para lectura crítica, como las [plantillas CASPe](#), o una lista de verificación, como la [CONSORT](#), para no dejaros ninguno de estos puntos sin considerar. Eso sí, todo lo que hemos hablado se refiere a ensayos clínicos aleatorizados y controlados, ¿Y qué pasa si se trata de ensayos no aleatorizados o de otra clase de estudios cuasiexperimentales?. Pues para eso se siguen otra serie de normas, como pueden ser las de la declaración [TREND](#). Pero esa es otra historia...
