CAPÍTULO 5

Lectura crítica de estudios de tratamiento. Ensayos clínicos aleatorios

Juan Bautista Cabello López ■ Eduardo López Briz
José Ignacio Pijoan Zubizarreta

OBJETIVOS DEL CAPÍTULO

- Definir ensayo clínico aleatorio (ECA) y su importancia para evaluar intervenciones.
- Clarificar los términos y conceptos clave para juzgar la validez de un ECA.
- Describir las características metodológicas capaces de influir sobre la validez e identificar sus consecuencias sobre diferentes dominios del estudio: riesgo de sesgo.
- Comprender las estrategias lectoras para identificar esos riesgos de sesgo.
- Interpretar los resultados de los ECA en términos del efecto de la intervención y de su relevancia.

Introducción

La mayoría de las preguntas que se formulan en la clínica corresponden a preguntas sobre la efectividad de los tratamientos o intervenciones, sean con finalidad preventiva o terapéutica (v. capítulo 3). Para obtener respuestas a este tipo de preguntas el ensayo clínico aleatorio (ECA) es considerado tradicionalmente el diseño de investigación clínica de referencia. Por tanto, saber leer críticamente un ECA es de una importancia capital para la práctica basada en la evidencia.

El ECA es un diseño prospectivo de investigación clínica (en personas con un problema específico de salud o en riesgo definido de desarrollarlo) en el que se evalúa el efecto de al menos dos intervenciones alternativas por medio de la asignación explícita (por un mecanismo aleatorio) a una de ellas de cada participante y la comparación de los desenlaces obtenidos en cada grupo generado.

Este diseño incorpora una herramienta específica (la aleatorización) que potencia la validez interna de sus resultados y justifica el alto nivel de credibilidad que se otorga a los mismos. En efecto, la asignación aleatoria evita los subjetivismos conscientes o inconscientes en la creación de los grupos de comparación (sesgos de selección) y genera grupos cuyo pronóstico inicial es similar, creando, de ese modo, el escenario ideal para comparar (sin confusión) el efecto de dos o más intervenciones. En suma, es un diseño que ofrece, junto a una muy alta solvencia epistemológica, una apariencia de comparación sencilla e intuitiva.

Hay que señalar que los resultados de las intervenciones son generalmente múltiples (unos positivos y otros negativos), y en tal sentido los ECA proporcionan información sobre ambos y por tanto datos para estimar el balance beneficio/riesgo.

Sin embargo, los ECA están, fundamentalmente, diseñados para responder a cuestiones de eficacia o efectividad, y aunque analizan la seguridad en el horizonte temporal del ensayo, pueden no detectar efectos adversos relevantes que sean poco frecuentes o tardíos. Estas «otras» consecuencias de las intervenciones se identifican *a posteriori* mediante estudios observacionales (estudios de cohortes o casos y controles) o por el uso ulterior en la rutina o registros asistenciales (1). Finalmente, tras enfatizar la importancia de estos diseños para este tipo de preguntas, conviene señalar, también, que hay preguntas sobre la efectividad de algunas intervenciones para las que este diseño de estudio resulta absolutamente superfluo (2).

En realidad, el ECA es toda una familia de diseños que comparten lo esencial: una estrategia de comparación y la aleatorización. Aunque mencionaremos algunos de los subtipos de esa familia de ECA, en este capítulo nos referiremos, por defecto, al diseño típico y más común: el ECA paralelo. En cuanto a las convenciones de escritura de ECA, cuya importancia mencionábamos en capítulos previos, para este diseño paralelo la convención al uso es CONSORT (***Con****solidated* ***s****tandards* ***o****f* ***r****eporting* ***t****rials,* http://www.consort-statement.org/), y para las diferentes variantes de diseño, las respectivas Extensiones de CONSORT.

Escenario

Ves en tu consulta a Manuela, una mujer de 46 años sin antecedentes de alergia a betalactámicos, historia de hipertensión arterial de 5 años de evolución que controla con inhibidores de la ARA II, mantiene reglas regulares y vida sexual activa y realiza ejercicio físico habitualmente. Ella tiene historia de infecciones urinarias de repetición desde hace 3 años, que han sido tratadas con cotrimoxazol forte. Hace 2 años, aconsejada por una amiga, siguió un régimen rico en frutos del bosque con escaso resultado, y más recientemente intensificó ese plan tomando preparados con extracto de arándanos y otros frutos del bosque con análogos resultados. Consulta actualmente porque ha recibido un comentario en una red social de que la mayoría de las infecciones de orina van muy bien con ibuprofeno.

Te pregunta tu opinión y le indicas que revisarás el asunto y en próxima consulta le informarás.

Buscas en las guías de práctica sin resultados y vas a las bases de datos de estudios primarios, donde encuentras este estudio.

Vik I, Bollestad M, Grude N, Bærheim A, Damsgaard E, Neumark T, et al. Ibuprofen versus pivmecillinam for uncomplicated urinary tract infection in women—A double-blind, randomized non-inferiority trial. PLoS Med 2018;15(5):e1002569.

Te preguntas:

1. ¿Es efectivo el ibuprofeno frente a pivmecilinam para prevenir infecciones urinarias en mujeres?
2. ¿Prescribirías ibuprofeno a Manuela?

Puntos clave de la lectura crítica de un ensayo clínico aleatorio (ECA)

La «calidad» de un estudio de un ensayo clínico es un concepto o constructo complejo cuya definición incluye diferentes elementos (o componentes): relevancia de la pregunta de investigación, adecuación y eficiencia del diseño, corrección bioética, excelencia en la ejecución y el análisis de los datos y rigor en la interpretación de los mismos, validez de sus resultados, corrección en la escritura, adecuación de la autoría, etc. Esta es, obviamente, una definición muy amplia que puede tener interés en algunos ámbitos, pero para ámbitos clínicos importan especialmente tres de esos componentes citados.

El primero es la pertinencia clínica, entendiendo por tal que se trate de preguntas de investigación que afecten a problemas de la clínica y, sobre todo, que incorporen desenlaces de investigación útiles para la toma de decisión clínica.

El segundo es la «corrección metodológica» del ensayo o en qué medida el diseño, la conducción y el análisis minimizan los sesgos (de selección, medición y confusión) en la estimación de efecto de la intervención. Es decir, ¿cuál es el ***riesgo de sesgo*** del estudio? (3), o por decirlo de modo más práctico, ¿en qué medida nos vamos a creer los resultados del estudio?

El tercero es la aplicabilidad o transferibilidad del resultado a un paciente concreto (o a un grupo de pacientes) considerando los otros elementos que influyen en la aplicación de esa evidencia (balance riesgos/beneficios, disponibilidad, valores del paciente, costes, etc.).

Abordaremos secuencialmente los tres aspectos citados, aunque en este capítulo pondremos el acento en los dos primeros componentes y trataremos sobre aplicabilidad posteriormente, a propósito del establecimiento de recomendaciones (v. capítulo 18).

LAS PREGUNTAS DEL ENSAYO (PERTINENCIA Y UTILIDAD DECISIONAL)

En el capítulo 3 señalábamos cómo se construyen las preguntas clínicas en formato estructurado o PICO (paciente, intervención, comparación y desenlace). Este formato contiene implícita la arquitectura del estudio: en una población definida evaluaremos paralelamente los desenlaces de una intervención comparada con otra definiendo un horizonte temporal. Se trata de un estudio longitudinal, similar a un estudio de cohortes del que se diferencia en que la asignación a uno u otro grupo se realizará de modo aleatorio, acorde con su carácter de experimento (fig. 5.1). Analicemos por separado los elementos de esa pregunta.

Población

Cuando formulábamos preguntas clínicas hablábamos de pacientes (P). Ahora, en investigación, la P corresponderá a la «población de estudio» y se define como el subgrupo de la población que reúne los criterios de inclusión y carece de los de exclusión.

Idealmente, esta población debe parecerse a los pacientes en quienes se piensa aplicar la intervención si fuera efectiva, y no es preciso ningún muestreo representativo. Frecuentemente, por razones de índole práctica o regulatoria se selecciona un subgrupo en el que es más probable observar el desenlace investigado y en ocasiones ingresan «a prueba» *(run-in)* en el estudio y solo son reclutados definitivamente si cumplen ciertos criterios. Por todo ello la población estudiada suele reflejar solo una parte del espectro habitual de la enfermedad o condición.

Concierne a la prudencia clínica valorar las diferencias entre las poblaciones de estudio y el paciente concreto a quien debe aplicarse el resultado. En realidad, casi nunca encontraremos un ensayo con una población exactamente igual a nuestro paciente; por ello la pregunta sería si nues-

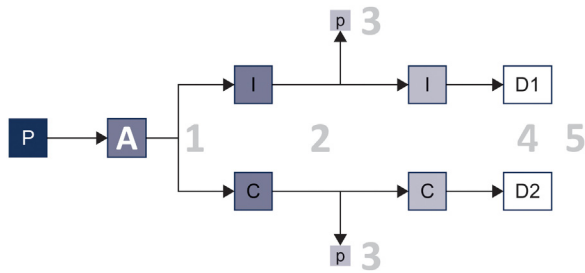

Figura 5.1 Esquema de un ensayo clínico aleatorio. Los números representan los dominios del estudio donde hay que valorar el riesgo de sesgo: 1, problemas en la aleatorización; 2, desviaciones de las intervenciones previstas; 3, pérdidas; 4, medición de los desenlaces; 5, reporte selectivo de los resultados. A, aleatorización; C, grupo de control; D1, desenlace en grupo de intervención; D2, desenlace en grupo de control; I, grupo de intervención; p, perdidos; P, población de estudio.

tro paciente es tan distinto como para no poder aplicarle el resultado del ensayo. Este proceso de aplicación es realmente, en sentido aristotélico, una cuestión prudencial.

Intervención

Se trata de la acción cuyo efecto se pretende estudiar y que puede ser de diferentes tipos: fármacos, grupos de fármacos, dispositivos, intervenciones quirúrgicas, fisioterápicas, psicológicas, estrategias de manejo clínico o estrategias de organización, intervenciones complejas, terapias combinadas y un largo etcétera.

En cualquiera de los casos la intervención (y la comparación) debe ser estandarizada y descrita con el detalle suficiente como para poder ser aplicada a la clínica, reproducida en investigación o incorporada en investigación de síntesis. La estandarización es relativamente fácil cuando hablamos de fármacos, pero las intervenciones más complejas exigen definiciones más elaboradas y contextualizadas. Otras veces las intervenciones son críticamente dependientes de las habilidades específicas de las personas que las realizan (por ejemplo: cirugía, hemodinámica, manipulaciones vertebrales, psicoterapia, etc.) y la estandarización resulta más complicada. En tales casos el control de las curvas de aprendizaje o la utilización de diseños especiales llamados «diseños de habilidad» *(expertise design)* son alternativas disponibles.

En otras ocasiones la intervención no se puede realizar sobre un individuo, pues hay que considerar su integración en una organización o colectividad. Por ejemplo, al estudiar estrategias docentes entre grupos aleatorios de residentes del mismo hospital la contaminación entre grupos es segura, o estudiar dos intervenciones dietéticas entre escolares de la misma escuela plantea dificultades obvias. En esos casos puede ser de utilidad el diseño de ensayo en *clusters* o conglomerados, cuya complejidad excede los objetivos del capítulo. Finalmente, cuando el efecto de la intervención es reversible en un plazo corto y se aplica a una enfermedad crónica estable es posible probar sucesivamente intervención y comparación en la misma población tras un período de lavado: se trata de los diseños cruzados *(cross-over trials)*. Este diseño puede ser llevado al extremo en algunas circunstancias (paciente no adecuadamente representado en la población de estudio de los ensayos disponibles u otras situaciones en las que los resultados de los ensayos no sean directamente aplicables a nuestro paciente) aplicándolo a un único paciente (ensayos de $n = 1$); en estos casos, la conexión entre la investigación clínica y su aplicación es máxima y puede ser el único método de valorar la mejor intervención para un individuo concreto (4).

Comparación

En este aspecto está, sin duda, la clave práctica y ética de los ECA. Desde el punto de vista del clínico práctico solo tiene sentido comparar nuevas intervenciones con intervenciones con efectos ya probados, o al menos que sean los tratamientos usuales; no usar tratamientos probados sería maleficencia y además esa comparación reproduce el posible dilema decisional real (tratamiento nuevo frente a tratamiento usual).

Desde el punto de vista ético, para proponer a un paciente la participación en un ensayo debe existir un equilibrio entre los posibles beneficios y riesgos esperables del nuevo tratamiento con los beneficios y riesgos de la intervención comparada (es la llamada *equipoise*). En realidad, este concepto refleja el punto de fricción entre dos dialécticas diferentes: la de práctica clínica y la de investigación clínica, y es, por tanto, un asunto crucial que condiciona el diseño en varios sentidos.

En primer lugar, exige un conocimiento explícito del estado del tratamiento para la condición clínica en cuestión (preferiblemente mediante una revisión sistemática). En segundo lugar, la existencia de tratamientos efectivos limita el uso de placebo como técnica de investigación y obliga a incluirlos en las comparaciones. En tercer lugar, la existencia de tratamientos efectivos condiciona, cada vez más frecuentemente, la elección de unos de diseño especiales conocidos como estudios de no inferioridad o de equivalencia. En cuarto lugar, hay ocasiones en las que se considera que ese balance entre riesgos y beneficios que llamamos *equipoise* puede cambiar durante el ensayo

(como consecuencia de él); en tales casos tiene interés realizar diseños secuenciales (que tampoco abordaremos) o programar análisis intermedios. Una posible consecuencia de ese cambio en el balance riesgo/beneficio es que sea preciso suspender el ensayo.

En realidad, las razones para terminar de forma precoz un ensayo son fundamentalmente tres: 1) el beneficio observado del tratamiento experimental es muy superior a lo esperado *a priori;* 2) el beneficio esperado de la nueva intervención, si existe, es inferior a lo esperado, poco relevante y es altamente improbable que el ensayo, en su diseño y dimensión original, sea capaz de detectar las diferencias esperadas (finalización por futilidad del efecto), y 3) los efectos adversos y la toxicidad del nuevo tratamiento son superiores o más graves de lo esperado.

Existen distintos procedimientos para realizar análisis repetidos de los datos sin afectar a la integridad estadística del análisis global, pero persiste un importante debate sobre cuándo se dispone de suficiente evidencia para considerar que la incertidumbre sobre el riesgo/beneficio no se mantiene, y ha de suspenderse el estudio. En cualquier caso, hay evidencias empíricas de que los ensayos finalizados prematuramente por detección de un beneficio superior al esperado suelen aumentar la incertidumbre en vez de disminuirla, ya sea por obtener estimaciones iniciales de beneficio muy optimistas que no se confirman en estudios posteriores o porque se centraron en variables subrogadas sin clara correspondencia con el desenlace clínico fundamental (5,6). Por ello la presencia de detención precoz del estudio debe ser mirada con cautela por el lector clínico.

Desenlaces

Pueden ser orientados a los pacientes u orientados a la enfermedad (generalmente desenlaces subrogados). El catálogo de desenlaces es tan amplio como la clínica: puede tratarse de condiciones clínicas objetivas como, por ejemplo, mortalidad, eventos clínicos como accidente vascular cerebral o infarto de miocardio o curación, pero también pueden ser síntomas, signos, percepciones, habilidades, calidad de vida, etc. En unos casos son valorados o medidos por médicos, sanitarios o cuidadores, y en otras ocasiones pueden ser evaluados y/o comunicados directamente por el propio paciente. Sea como fuere, será preciso evaluarlos con cuidadosa visión clinimétrica.

En otro sentido, la arquitectura del estudio permite comparar simultáneamente muchos desenlaces clínicos, y por ello existen en los ECA dos jerarquías de desenlaces. Una es la investigacional, en la cual los desenlaces (o variables de resultado) son clasificados en principal y secundarios según condicionen o no el diseño del estudio (tamaño muestral, sistemas de medición, otros elementos metodológicos, etc.). La otra jerarquía es la de la importancia clínica, que depende de su influencia decisional y que, como señalamos en el capítulo 3, podía ser de tres tipos: desenlaces críticos para la decisión, desenlaces importantes-no-críticos para la decisión y desenlaces no importantes.

Son dos visiones obligadas a coexistir, pero desde la perspectiva del lector hay que plantearse dos aspectos: el primero es si son estos los desenlaces que necesito para mi decisión clínica o para mi investigación (es decir, ¿es esta es mi pregunta?). El segundo es si están convenientemente comunicadas todas las variables relevantes o, al menos, ¿están todas las prometidas en el protocolo?

A veces un grupo de síntomas, signos o variables, consideradas en conjunto, reflejan mejor el estado de salud o el efecto del tratamiento que tomadas por separado; en esos casos hablamos de variables compuestas (un ejemplo de ellas es la **ACR50** del capítulo 3).

En otros casos el desenlace supone la terminación de la contribución de un paciente al ensayo (por ejemplo, se produce su muerte), o se produce el evento que estamos estudiando (infarto de miocardio, accidente vascular cerebral). Este tipo de desenlaces se denominan «punto final» *(end point)*. Un caso algo especial, frecuente en algunas áreas de investigación, es el de las variables punto final compuestas *(composite end point)*. Se trata de una variable que mide la ocurrencia de cualquiera de los eventos punto final que la constituyen y puede hacerlo como ocurrencia de alguno de ellos en el tiempo predefinido o como tiempo de ocurrencia hasta que acaece cualquiera de los elementos de la variable punto final compuesta. Por ejemplo, en un estudio sobre estatinas la variable «punto final compuesta» podría ser el tiempo hasta la ocurrencia de cualquiera de los eventos siguientes:

infarto de miocardio fatal, infarto no fatal, accidente vascular cerebral o accidente vascular periférico, o la ocurrencia de cualquiera de ellas en 1 año.

Este abordaje puede tener cierto sentido biológico y/o clínico en la medida en que informa sobre el progreso de la enfermedad vascular en su conjunto, y es usado frecuentemente porque aumenta la probabilidad del desenlace y con ello hace más eficiente el estudio al precisar menos tamaño de muestra. Pero a efectos de su lectura debe ser mirado con suma prudencia (7) y hemos de obtener información de cada variable por separado y en su conjunto. En todo caso, a efectos de decidir su importancia clínica puede haber dificultades, porque los diferentes componentes pueden ser heterogéneos cualitativa y/o cuantitativamente.

En resumen, el lector debe decidir si estas son sus preguntas, si las comparaciones son las adecuadas, si el diseño es pertinente, si los desenlaces son los realmente importantes para su paciente o para su investigación, si la pregunta está formulada en términos de superioridad o de no inferioridad y si la población es tan distinta de su caso (o de su población de interés) que no podrá aplicar sus resultados.

RIESGO DE SESGO DEL ESTUDIO

Desde una perspectiva «lectora», validez significa realmente explorar de modo concienzudo aquellas características metodológicas que son potencialmente capaces de producir errores sistemáticos (o sesgos) en la estimación del efecto (3). Sin embargo, un estudio determinado puede producir estimaciones correctas para algunos desenlaces y sesgadas para otros desenlaces del mismo estudio; por ejemplo, un estudio, no ciego, sobre el efecto de la administración de oxígeno frente a aire en el infarto agudo de miocardio puede estimar correctamente el efecto sobre la mortalidad, pero, al no ser ciego, estimar sesgadamente su efecto sobre el alivio del dolor. Por tanto, la repercusión sobre la validez del estudio de las características metodológicas debe realizarse independientemente para cada desenlace. Este es un principio básico para la lectura de un ECA.

Esas características citadas son: la aleatorización, la ocultación de la secuencia de aleatorización, el cegamiento en sus múltiples variantes, las pérdidas de pacientes del estudio, la medición del desenlace y el reporte selectivo de los desenlaces (fig. 5.2).

Tales características pueden afectar de diferente modo a las diversas partes del estudio, o lo que llamamos «dominios» de estudio que se muestran en la figura 5.1 y se describen a continuación.

En esencia, un ensayo clínico aleatorio es una estrategia de comparación que se basa en cinco pilares: el primero es la construcción de grupos que se van a comparar tan iguales como sea posible, mediante la aleatorización; los problemas con la aleatorización pueden afectar a esa construcción de grupos y ser fuente de sesgo (Dominio 1). El segundo es el mantenimiento de la comparabilidad de los grupos a lo largo de todo el estudio, por ello las desviaciones de las intervenciones previstas, sean derivadas de la asignación efectiva a los grupos o de problemas en la adhesión de los pacientes a los

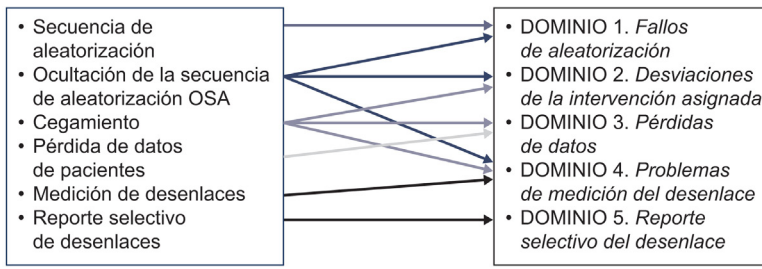

Figura 5.2 Características metodológicas que se han de explorar en los ensayos clínicos aleatorios y dominios del estudio sobre los que puede producir sesgos cada característica. Las características son usadas en el instrumento de Cochrane RoB1 (22), mientras que los dominios son usados en el Cochrane RoB2 (23).

TABLA 5.1 ■ **Principales formas de aleatorización**

Tipo de aleatorización	Procedimiento
Aleatorización simple	Cada paciente tiene una probabilidad (habitualmente la misma) de ser asignado a uno u otro grupo, y ello no puede ser predicho. Los métodos son moneda, tabla de números aleatorios o generación por ordenador (números pseudoaleatorios)
Aleatorización restrictiva	Se impone alguna restricción al proceso de aleatorización (p. ej., bloques) para asegurar el equilibrio de pacientes entre los grupos
Aleatorización balanceada por covariables	Frecuentemente se desea similar número en cada grupo, pero también asegurarse de que los «factores pronósticos» importantes sean similares en ambos grupos. Existen varios mecanismos: *estratificación, estratificación + bloques o algoritmos de minimización*
Aleatorización adaptativa según respuesta	La probabilidad de asignación al tratamiento depende de las respuestas previas al mismo

Adaptado de McPherson 2012 (8).

grupos respectivos, pueden alterar la comparabilidad de los grupos (Dominio 2). Esas desviaciones de las intervenciones (con cambios, abandonos, retiradas, etc.) implica que habrá diversos modos de analizar estos datos y según el modo habrá más o menos riesgo de sesgo para la estimación del efecto. El tercero es que, dado que el estudio tiene dimensión temporal, muy probablemente se perderán algunos pacientes (y sus datos) y esas pérdidas pueden, también, amenazar la comparación (Dominio 3). En cuarto lugar, la medición de los desenlaces debe ser adecuada y equilibrada; problemas en este apartado pueden, asimismo, amenazar la comparación (Dominio 4). En quinto y último lugar, el reporte selectivo de los resultados puede romper el equilibrio de la comparación y ser igualmente fuente de sesgo (Dominio 5).

A continuación, describimos esas características metodológicas mencionadas y el impacto que cada una de ellas puede producir sobre los diferentes dominios del estudio.

Aleatorización

Consiste en la asignación de cada paciente a cada grupo en función de una secuencia aleatoria generada de diversos modos posibles (tabla 5.1) (8). Esta es la manera que tenemos de intentar que los dos grupos sean similares respecto de las variables que creemos importantes, pero también de otras variables desconocidas. En teoría debe controlar el sesgo de selección y el sesgo de confusión. Conceptualmente hay dos elementos distintos en la aleatorización, uno es la generación de la secuencia de aleatorización y otro, más pragmático, es cómo se realiza el procedimiento de asignación y especialmente si la secuencia se mantiene oculta para quien hace el reclutamiento (ocultación de la secuencia de aleatorización, OSA).

Secuencia de aleatorización. Puede generarse de modo simple a través de una tabla o sistema de números aleatorios (hasta con una moneda si hay solo dos grupos a comparar). El problema de las secuencias generadas así es el desequilibrio de efectivos entre grupos, que es especialmente frecuente para ensayos pequeños. El modo de resolver este problema es haciendo bloques de permutaciones (aleatorización restrictiva por bloques); de ese modo la máxima diferencia posible entre grupos en el número de individuos será igual a la mitad del tamaño del bloque. El segundo problema posible es la distribución desigual en los grupos de los factores pronósticos importantes, lo cual puede afectar al control del sesgo de confusión. Este problema suele abordarse mediante la estratificación (con o sin bloques por estrato), o mediante algoritmos de minimización (9)

que resuelven el problema de manejar muchos estratos y cuyo uso se ha popularizado en los últimos años (3).

En cualquiera de los casos la aleatorización genera una tendencia a la similitud entre los grupos, pero solo a largo plazo (con tamaños muestrales grandes). Sin embargo, pese a las precauciones, el azar puede producir desequilibrios en los factores pronósticos. Por tanto, la efectividad de la aleatorización debe ser comprobada en todos los ensayos. El efecto de la aleatorización sobre las variables conocidas suele mostrarse en la tabla 1 de todos los ensayos (es la llamada tabla 1 de CONSORT) cuya exploración es obligada. En ella puede verse la distribución de características en ambos grupos y suele realizarse test de significación estadística, de dudoso valor. La presencia de diferencias importantes en la distribución de variables pronósticas en los grupos, aún sin significación estadística, o la presencia de diferencias significativas en muchas características de los pacientes puede sugerir que la randomización no ha sido eficaz.

Ocultación de la secuencia de aleatorización (OSA). El problema de la predictibilidad del grupo de tratamiento se debe a que saber a qué grupo se asignará el próximo paciente puede condicionar los comportamientos clínicos o de quien recluta (10). Imaginemos que sabemos que el próximo paciente reclutado será asignado al grupo experimental, y tenemos una cierta preferencia, consciente o no, por uno de los tratamientos (el tradicional). Supongamos que estamos ante un paciente con criterios límite de inclusión (especialmente grave) y creemos que en realidad le beneficiaría más el tratamiento clásico; es muy posible que forcemos la exclusión del paciente, y con ello estemos generando un sesgo de selección al favorecer al nuevo tratamiento. En resumen, conocer la secuencia puede generar sesgos de selección y eliminar en parte las ventajas de la aleatorización.

Esta OSA debe diferenciarse del cegamiento, que será comentado después. En efecto, se puede mantener la secuencia de aleatorización oculta y sin embargo tratarse de un ensayo en el que se compara un procedimiento quirúrgico con uno médico y por tanto es un ensayo abierto.

A efectos de lectura, la no comunicación explícita de la secuencia suele asociarse a otros déficits metodológicos. En cuanto a la OSA, hay consistente evidencia empírica (10-13) de que su ausencia provoca una sobreestimación del efecto de hasta el 40% y es, sin duda, una de las más importantes causas de sesgo en los ECA.

Encargar la aleatorización a la farmacia del centro, el uso de sobres opacos ordenados y numerados o, preferiblemente, usar una central de aleatorización son los métodos de afrontar el problema. Un aspecto final que se debe señalar es que el uso de bloques y la minimización pueden en algunos casos hacer predecible la siguiente asignación y con ello desocultar la secuencia de aleatorización; esto debe ser también considerado en la lectura.

En cuanto al dominio donde impactan tanto el desbalance en los grupos por efecto del fracaso en la aleatorización como la ausencia de OSA es en el DOMINIO 1, es decir, altera la construcción de grupos comparables. Adicionalmente, la ausencia de OSA puede hacer que personal del estudio conozca el grupo asignado al paciente, y ello puede afectar a los cuidados paralelos (DOMINIO 2) o a la evaluación del desenlace (DOMINIO 4).

Cegado (enmascaramiento)

Entendemos por cegado o enmascaramiento en un ECA el procedimiento por el cual se asegura que los participantes, los clínicos, los investigadores, los medidores de los desenlaces o los que analizan el estudio desconozcan qué intervención se administra a cada participante. La figura 5.3 muestra algunas definiciones relacionadas con los distintos tipos de cegado.

La ausencia de cegamiento en los pacientes puede producir disbalance en las intervenciones porque los grupos asignados a la intervención experimental suelen ser más proclives a tener otros comportamientos saludables, y la consciencia de la asignación a algunas intervenciones puede influir en la solicitud de cuidados adicionales (DOMINIO 2). Por otra parte, la propia dinámica del ensayo puede modificar el equilibrio de los grupos; por ejemplo, la consciencia de asignación al grupo

ECA abiertos *(open label)*
Ensayos en los que la naturaleza de las intervenciones asignadas a los participantes es conocida tanto por ellos como por los investigadores durante todo el ensayo.

ECA simple ciegos *(single blind)*
Ensayos en los que la naturaleza de las intervenciones asignadas a los participantes no es conocida por ellos, pero sí por los investigadores durante todo el ensayo.

ECA doble ciegos *(double blind)*
Ensayos en los que la naturaleza de las intervenciones asignadas a los participantes no es conocida ni por ellos ni por los investigadores a lo largo de todo el ensayo.

ECA triple ciegos *(triple blind)*
Ensayos en los que la naturaleza de las intervenciones asignadas a los participantes no es conocida ni por ellos, ni por los investigadores, ni por las personas que analizan los datos.

Doble enmascaramiento *(double dummy)*
Técnica para mantener el cegado cuando a los tratamientos no se les puede dar la misma apariencia. Se prepara un juego para el fármaco A (medicamento activo y placebo idénticos) y otro para el fármaco B (medicamento activo y placebo idénticos). Cada paciente recibirá A activo + B placebo o B activo + A placebo, de manera que se respeta la isoapariencia.

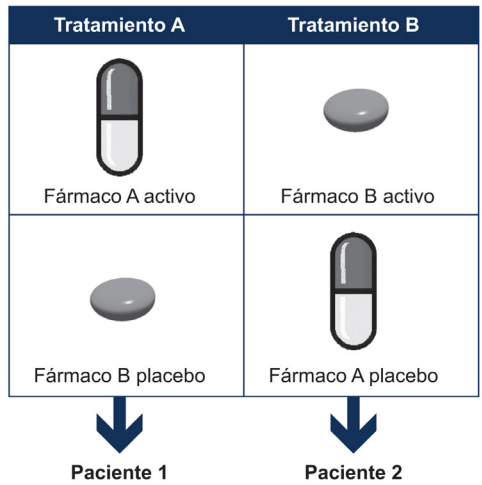

Figura 5.3 Algunas definiciones relacionadas con el cegado en los ensayos clínicos aleatorios. ECA, ensayo clínico aleatorio.

control en un ensayo abierto puede inducir al paciente a buscar otras intervenciones adicionales similares a la experimental (DOMINIO 2), o la ruptura de cegamiento por efectos colaterales puede propiciar el abandono del estudio por el paciente (DOMINIO 3). La ausencia de cegamiento puede también producir diferentes grados de adherencia a la intervención y/o pérdidas diferenciales en los grupos en aquellos casos en los que mantenerse en el ensayo (o en algunos de sus brazos) requiera de cierto entusiasmo del paciente (DOMINIO 3). Finalmente, la ausencia de cegamiento en el paciente puede a afectar a la aparición diferencial de síntomas subjetivos o a la medición de los mismos, especialmente si los desenlaces son medidos por el propio paciente (DOMINIO 4).

La ausencia de cegamiento en los clínicos y/o investigadores puede hacer que la administración de cuidados extraprotocolarios sea diferencial en uno u otro grupo (DOMINIO 2) y también puede provocar, en casos límite de efectos colaterales, retiradas del paciente por el clínico que serán diferenciales en los grupos (DOMINIO 3). Otro efecto de la ausencia de cegamiento es su impacto

sobre el proceso de medición de los desenlaces al introducir preconcepciones en el observador o en el procedimiento. El posible impacto del cegado sobre las mediciones depende del tipo de cegado, pero también del tipo de variable a medir (DOMINIO 4).

Hay evidencia empírica que demuestra que la ausencia de cegado puede sobreestimar el efecto en más del 20%, especialmente si los desenlaces son variables subjetivas o síntomas, aunque (11-13), probablemente por variabilidad del cegamiento y la complejidad de sus efectos, no hay consistencia total al respecto en todos los estudios (14). Es, por tanto, un tema que precisa más aproximaciones.

El impacto del enmascaramiento es, como vemos, diverso y complejo, por lo que, desde la perspectiva del lector, será preciso valorar cuidadosamente en cada estudio cómo los detalles concretos del cegado pueden influir sobre cada uno de los dominios señalados.

Datos perdidos de desenlaces

Como el ECA tiene un cierto horizonte temporal, es bastante verosímil que se produzcan pérdidas en el seguimiento de los pacientes por distintas causas: no podemos localizar al paciente, pérdidas simples, o no se puede medir el desenlace por agravamiento de su enfermedad, fallecimiento o algún problema concomitante; puede también ocurrir que el paciente no abandone el estudio, pero no se disponga de todos los datos requeridos. Estas pérdidas pueden en algunos casos producir riesgo de sesgo (DOMINIO 3).

La importancia depende, obviamente, del valor clínico del desenlace en cuestión y, sobre todo, de si la pérdida tiene (o puede tener) relación con el resultado de ese desenlace o/y con la pertenencia al grupo experimental (3). Podemos sospechar que las pérdidas dependen del desenlace cuando el número de ellas difiera en ambos grupos o en el caso de tiempo de fallo cuando los casos censurados difieran en ambos grupos. Hay que reflexionar en cada caso porque hay áreas de estudio en las que es conocido que la pérdida se relaciona con el valor del resultado; por ejemplo, en ensayos de intervenciones sobre esquizofrenia la pérdida suele asociarse a la ausencia de efecto del tratamiento.

Respecto del número para decidir si son pocas o muchas pérdidas, tradicionalmente se considera pocas por debajo del 5% para variables continuas y muchas por encima del 20%. Sin embargo, esos umbrales tienen frágil fundamento; la importancia depende del tipo de desenlace, menos importante en desenlaces continuos y más en dicotómicos (v.gr. mortalidad) y del riesgo basal de ocurrencia del desenlace o evento: si el riesgo es muy bajo, los efectivos esperables serán pocos en ambos grupos y en tal caso incluso pérdidas exiguas pueden penalizar mucho la estabilidad de la estimación del efecto.

El mecanismo de las pérdidas es frecuentemente descrito en el informe del estudio. Para su detección pueden ser de ayuda las normas de CONSORT, y en particular el flujograma, que permite ver lo acontecido en el tiempo a todos los pacientes seleccionados y aleatorizados. También puede ser útil que la tabla 1 recomendada por CONSORT, que presenta las principales características basales de los pacientes en cada grupo de tratamiento según la aleatorización inicial, se expanda presentando para cada grupo las características de los pacientes que finalmente son incluidos en el análisis seguido de las características de aquellos excluidos. De esta forma el lector podrá valorar hasta qué punto hay diferencias en pérdidas entre ambos grupos y su posible repercusión (15).

Una vez detectada la presencia y calibrada la importancia de este sesgo, existen diferentes estrategias a la hora de releer o reanalizar los datos. La más intuitiva es desarrollar ciertos escenarios hipotéticos o simulaciones para asignar valores a los datos faltantes y evaluar hasta qué punto cambian los resultados y las conclusiones fundamentales del ensayo. Es el llamado análisis de sensibilidad, que admite dos escenarios extremos: el «análisis en el peor de los casos», en el que se calculan los estimadores del efecto si todos los participantes perdidos del grupo experimental tuvieran el evento negativo y los perdidos del grupo control no lo tuvieran, y el «análisis en el mejor de los casos», en el que procederíamos justo al revés. De este modo veremos cuál es la sensibilidad de nuestra estimación a los posibles cambios en las pérdidas. Entre estos dos escenarios extremos se pueden plantear otros intermedios, más o menos plausibles, en función del problema clínico

de estudio, que pueden enriquecer la perspectiva sobre la robustez de los resultados y las posibles relaciones entre las pérdidas y los resultados.

Otra alternativa es realizar análisis ajustados por una o más covariables que muestren desequilibrios entre los grupos que se analizan o utilizar determinadas técnicas estadísticas que «imputan» o asignan valores a los datos faltantes. Existe un número creciente de metodologías analíticas, algunas muy sofisticadas, para aquellas situaciones en las que hay datos faltantes, pero no debemos perder la perspectiva, pues bajo la maquinaria de los modelos estadísticos subyacen suposiciones teóricas de difícil comprobación sobre los mecanismos que han originado la ausencia de los datos (16).

Medición adecuada de desenlaces

El catálogo de desenlaces medibles es sumamente diverso, como mencionábamos en el apartado «Desenlaces», y los problemas en la medición se denominan error de medición para las variables continuas, mala clasificación para las dicotómicas y no comprobación o verificación para los eventos. En unos casos son valorados o medidos por médicos, sanitarios o cuidadores, o también por comités externos independientes del estudio, y en otras ocasiones pueden ser evaluados y/o comunicados directamente por el propio paciente.

Los problemas de medición en el ensayo pueden afectar por igual a ambos grupos (no diferenciales) o afectar de modo diferente a los grupos (diferenciales), en cuyo caso el riesgo de sesgo sobre este dominio aumenta.

Este asunto cabe abordarlo de modo tradicional considerando el hecho que se va a medir en el sujeto *(input),* el método de medición y el papel del observador. Lo más fácil es juzgar si el procedimiento de medición es adecuado o no (por sus características clinimétricas, factibilidad, momento, tolerabilidad, etc.). El segundo aspecto del juicio es si los efectos que se van a medir (el *input* de la medición) pueden estar influenciados por el diseño del estudio, como puede ocurrir con síntomas subjetivos si el paciente conoce el grupo al que fue asignado o el ciego fue roto.

El tercer aspecto es si el observador que realiza la medición es ciego respecto de la intervención aplicada o si se han producido problemas en la OSA. Esta medición es particularmente difícil cuando el desenlace implica algún juicio o decisión clínica, y en esos casos suele usarse un comité externo para minimizar el riesgo de mala clasificación diferencial. Naturalmente, en los casos en los que el paciente es quien desarrolla el síntoma y realiza la medición, la ausencia de ciego afecta al *input* de la medición y al observador.

En cualquiera de los casos, estos problemas de medición afectarán al DOMINIO 4 y procede, como citábamos, una aproximación clinimétrica y un análisis en cada caso y para cada desenlace del impacto sobre la estimación del efecto.

Comunicación selectiva de desenlaces

Es un subtipo de sesgo de comunicación que consiste en la selección de un subgrupo de desenlaces (o variables) para la publicación del ensayo. Cuando se compara las publicaciones del ensayo con el protocolo del mismo se observa que en el 62% de los ensayos al menos un desenlace ha sido cambiado, introducido u omitido (17,18).

La comunicación selectiva de desenlaces puede adoptar formas muy diversas: omitir un desenlace, o publicarlo con insuficiente detalle para su aplicación, omitir parte de un desenlace compuesto, etc. A veces lo que cambia es la importancia asignada a una variable. Generalmente ocurre cuando se presenta como variable de desenlace principal una variable que en su momento se definió como secundaria.

Hay evidencias de que en diseños paralelos se publican solo el 50% de los desenlaces no significativos frente al 72% de los significativos, lo que supone un *odds ratio* (OR) de 2,4. Esa asimetría se mantiene tanto para los desenlaces de daño o perjuicio (OR 1,9 [IC 95% 1,1-3,5]) como para los de efectividad (OR 2,0 [IC 95% 1,6-2-7]) (18). Al leer el ensayo, esta comunicación selectiva puede hacer que veamos más fácilmente los efectos positivos y tengamos una sensación de beneficio aparente; por otra parte, plantea problemas adicionales para las revisiones sistemáticas de ECA.

La manera más simple de explorarlo es comparar el listado de desenlaces en la sección de material y métodos con los que luego son comunicados en resultados y tablas. Otro modo, más interesante y eficaz, es comparar la publicación del ensayo con el protocolo previamente publicado. A este respecto, los registros de ensayos, comentados en el capítulo 2, son de gran utilidad porque permiten comprobar y contrastar las variables previstas en el protocolo. Aunque en ensayos antiguos pueden no estar disponibles, en la actualidad no es posible publicar un ensayo no registrado, debido a que es requisito exigido por el Comité Internacional de Editores de Revistas Médicas (ICMJE).

Entonces ¿cómo analizamos los datos?

Las desviaciones de la intervención prevista, sea por problemas en la asignación o por problemas en la adherencia a la intervención (DOMINIO 2), plantean la cuestión de cómo analizar ese estudio en el que se han producido esos movimientos en los grupos en diferentes sentidos.

El aspecto esencial del análisis está relacionado con ¿cuál es el efecto de interés del estudio? En unos casos, el interés es estudiar el efecto de la asignación a la intervención (de la intención de tratar); esta es la aproximación pertinente si se desea conocer si debe aplicarse la intervención o no a una población determinada o en un sistema de salud. En otros casos, lo que interesa es estudiar el efecto de adherirse, es decir, al recibir realmente la intervención (19-21). Esta aproximación sería más adecuada para informar una decisión en un paciente concreto. En el primero de los casos hablamos de análisis por intención de tratar (AIT), y en el otro, de análisis por protocolo (APP) o de los tratados (AT) (19-21).

El **análisis por intención de tratar** consiste que cada paciente es analizado en el grupo al que fue aleatoriamente asignado con independencia del tratamiento que finalmente recibió o de otras circunstancias, e incluye a todos los pacientes en el análisis, lo que implica medir, al menos teóricamente, el desenlace de todos. La expresión clásica para enunciarlo es (una vez aleatorizado, siempre analizado; *once randomized, always analyzed*). Esta aproximación se centra en la población aleatorizada (P-AIT), que es un subgrupo de la población de estudio definida en el ensayo y respeta la aleatorización y sus efectos sobre el equilibrio de factores en los grupos.

Una variante de este es el **AIT modificado** (AITm), en el que se excluye del análisis a la subpoblación aleatorizada que tiene datos de desenlaces perdidos. En ocasiones se excluye a participantes que nunca iniciaron el tratamiento, que tras iniciarlo no han acudido a ninguna visita y por tanto no han aportado datos sobre los desenlaces o que tras la aleatorización eran no elegibles (por error en la selección o por problemas sobrevenidos). El AIT modificado no ha sido claramente definido y es usado con diferentes sentidos, por lo que en ocasiones se convierte en un instrumento para la manipulación de los datos.

El ***análisis por protocolo*** (APP) estricto consiste en que se incluye en el análisis solo a los sujetos que siguieron estrictamente el protocolo del estudio, y en consecuencia se excluye a los participantes que no recibieron su intervención asignada o se desviaron del protocolo. Es por tanto una subpoblación de P-AIT cuyos grupos son escindidos de los iniciales, y a diferencia de los grupos del AIT no son estrictamente comparables porque pueden ser afectados por sesgos de selección en el caso de que factores pronósticos estén relacionados con la adherencia al tratamiento.

Una forma algo distinta es hacer el **análisis de los tratados** (AT, *as treated*), en la cual los pacientes son analizados en el grupo de la intervención que realmente recibieron y completaron, independientemente de si fueron aleatorizados a otro tratamiento. Este es otro subgrupo distinto de la P-AIT con grupos reconstruidos en los que puede existir alto riesgo de sesgo si las razones por las que se pasan de grupo se asocian a factores pronósticos.

Respecto de cuál es entonces el método más adecuado, como señalábamos al principio, depende del interés de la pregunta del estudio.

Las ventajas que ofrece el de AIT son: que mantiene la aleatorización y por tanto controla la confusión y mantiene la comparabilidad, y en cierto modo se aproxima a lo que ocurrirá en general (los pacientes dejan de tomar el tratamiento y abandonan el estudio, etc.). Por otra parte,

es científicamente conservador en el sentido de que su uso produce sesgo hacia el no efecto (hacia la hipótesis nula), lo cual es aceptable. Por ello ha sido sugerido tradicionalmente como el modo más adecuado de análisis. Sin embargo, es obvio que el AIT implica ignorar deliberadamente todas las circunstancias y vicisitudes del estudio que sean posteriores a la randomización y que, por lo demás, son habituales en estos estudios clínicos.

Ese sesgo hacia la hipótesis nula no plantea mucho problema en los estudios de superioridad: estimamos menos efecto del real, es decir, sesga en contra del investigador; sin embargo, en estudios de no inferioridad, en los que formulación de la hipótesis es diferente, la dirección del sesgo será la de aparecer como un efecto inferior (menor efecto) cuando realmente es «no inferior».

El APP, como señalábamos, rompe el equilibrio de la comparación y transforma el estudio en algo más cercano a lo observacional, es decir, tiene más riesgo de sesgo. Sin embargo, es claro que responde a otra pregunta diferente. Por ello, para considerar el efecto real de la adherencia a la intervención o para ensayos pragmáticos, es adecuado el uso de APP, pero debe ser enunciado *a priori,* e idealmente asociado a una previsión y una definición razonable de la adherencia al protocolo y un control de sesgos mediante el uso de los instrumentos estadísticos de ajuste desarrollados al efecto (20,21).

Finalmente, para el caso de los ensayos de no inferioridad, lo aconsejable es usar los dos análisis e interpretar el APP como un análisis de sensibilidad.

Instrumentos para medir el riesgo de sesgo

Hasta aquí hemos hecho una aproximación argumental a la valoración crítica del ensayo a través del riesgo de sesgo. Este es un ejercicio de autonomía y empoderamiento clínico y se presta para propiciar el debate y la deliberación y con ello una mejor aplicación y un mejor aprendizaje. Pero cuando los ECA forman parte de un estudio de síntesis (revisión sistemática) es preciso que el proceso de medición del riesgo de sesgo sea consistente. Por ello se han construido instrumentos de evaluación del riesgo para esos estudios: actualmente se usa el Cochrane Risk of Bias 1 (RoB1) (22), que realiza para cada desenlace un juicio en tres categorías (bajo riesgo, dudoso, alto riesgo) sobre cada una de las características de los ensayos capaces de producir sesgo. Ese instrumento se ha perfeccionado recientemente (23) en el Cochrane (RoB2) con algunos cambios disruptivos: actualmente centra su interés y realiza sus juicios (bajo riesgo, dudoso, alto riesgo) sobre los dominios de estudio mencionados que guardan cierta correspondencia con las características (v. fig. 5.2) y para cada dominio del estudio incorpora un algoritmo de ayuda que, guiado por preguntas específicas, conduce a una de las tres categorías. Dado que ambos instrumentos van a coexistir durante un tiempo en las revisiones sistemáticas, hemos tratado de mostrar en este libro esa correspondencia entre características y dominios típica de cada uno de los instrumentos que puede ver el lector en el futuro.

RESULTADOS

Los ECA se llevan a cabo para determinar si una determinada intervención es efectiva y segura o si proporciona alguna ventaja en términos de riesgo/beneficio sobre una intervención de referencia. La decisión se basará en el análisis comparativo de los resultados obtenidos en cada grupo de intervención. Todos los elementos de calidad en el diseño y ejecución del ensayo comentados hasta ahora tienen como objetivo que esta comparación sea equilibrada (no artefactada).

Es importante determinar primero cuál es la escala de medida de la variable de desenlace principal, pues esta a su vez condiciona la técnica de análisis estadístico y la forma de presentación de los resultados. Si la variable principal se mide en una escala continua (por ejemplo, el nivel sanguíneo de un parámetro bioquímico o la puntuación de calidad de vida relacionada con la salud medida con el cuestionario SF-36), la forma habitual de expresar el resultado sería proporcionar la diferencia media entre el resultado observado en el grupo experimental y el del grupo de referencia, añadiendo información sobre la precisión de esta estimación en forma de error estándar de la misma o suministrando su intervalo de confianza. Generalmente se aconseja utilizar variables de gran

relevancia clínica, y entre ellas destacamos aquellas de tipo dicotómico o binario (SÍ/NO), pues se acomodan bien a la forma de trabajo del profesional asistencial (tiene/no tiene este problema, le trato/no le trato, se cura/no se cura, etc.) En este caso, deberemos siempre buscar los números «crudos» (es decir, en cuántos pacientes se evaluó el desenlace y en cuántos se detectó el desenlace de interés). De esta forma se clarificará si se analizó a todos los pacientes aleatorizados o a un subgrupo determinado. Además, con sencillas herramientas de cálculo epidemiológicas, el clínico entrenado y con ganas podrá fácilmente obtener y valorar a partir de ellos algunos estimadores del efecto quizá no presentados en el artículo y que pueden ser de utilidad interpretativa.

Con este tipo de variables dicotómicas podremos encontrar los siguientes descriptores de los resultados:

- El *Riesgo* o probabilidad de desarrollar el desenlace de interés (por ejemplo, curación) en el grupo de intervención experimental o el del correspondiente grupo de referencia; este riesgo podrá ser presentado en forma de una proporción (número decimal entre 0 y 1) o de un porcentaje (en escala de 0 a 100). Aunque esta no es *per se* una medida comparativa de resultado, es la base para la mayoría de las que describimos a continuación.
- La *Diferencia de Riesgos* (DR), también llamada *Reducción Absoluta del Riesgo* (RAR), que refleja la diferencia en la probabilidad de ocurrencia del desenlace entre los grupos. Una reducción de 0 equivale a igualdad de riesgo en los grupos y por tanto a la ausencia de efectos diferentes de las intervenciones que se comparan. Si se obtiene un resultado distinto de 0, habrá que valorar la magnitud y el sentido de la diferencia.
- El llamado *Número Necesario de Pacientes a Tratar* (NNT). Es el inverso de la DR (RAR) y nos informa del efecto de una intervención calculando cuántos pacientes deberían cambiar su tratamiento y recibir el tratamiento experimental EN VEZ del de referencia para (en el tiempo de seguimiento utilizado en el ensayo) conseguir un desenlace de interés adicional al que se obtendría si recibieran la intervención de referencia. Es fácil determinar que el inverso de 0 es infinito (una magnitud inespecífica) y por lo tanto este NNT no estimable sería el que nos hablaría de la no diferencia de efectos. El NNT es un estimador de la efectividad de la intervención, entendida como consecución de más desenlaces favorables o reducción de los desfavorables. Cuando queremos referirnos a los sucesos desfavorables, y específicamente cuando nos referimos a efectos indeseados de una intervención (toxicidad, efectos secundarios, etc.), se utiliza el llamado Número Necesario de Pacientes para causar Daño (NND, NNH en inglés), que informa sobre cuántos pacientes tendrían que recibir el tratamiento experimental en vez del de referencia para que observemos un daño o suceso desfavorable (generalmente un efecto tóxico grave) adicional a los que se observarían con el tratamiento de referencia o control. De esta forma, la relación NNT/NND nos permite una aproximación al balance beneficio/riesgo entre las intervenciones comparadas.
- El *Riesgo Relativo* (RR) de desarrollar el desenlace de interés en el grupo experimental RESPECTO del grupo de referencia. Si el RR es 1 asumimos que la probabilidad (riesgo) del desenlace es igual en los grupos y por tanto que NO hay un efecto diferente de una intervención respecto a la otra. Cifras superiores al 1 hablan de mayor riesgo del desenlace en el grupo experimental y cifras inferiores de menor probabilidad en dicho grupo. Existen otras dos medidas con una interpretación similar al RR: a) el *odds ratio* (OR) en el que no comparamos las probabilidades en la forma de manejo habitual en nuestro medio sino en una escala diferente (comparamos los *odds*); aunque en el contexto de un ensayo es más coherente la utilización del riesgo y sus estimadores derivados, el OR se utiliza básicamente porque es la forma de obtener estimaciones del efecto en modelos de regresión multivariable que tienen en cuenta el efecto añadido de otros factores y se usan frecuentemente en los análisis ajustados que hemos comentado anteriormente.
- El *Hazard Ratio* (HR) o razón de «riesgos», cuando la técnica estadística utilizada ha sido el análisis de supervivencia (generalmente, el llamado modelo de regresión de Cox). En estos

casos el desenlace de interés ha sido el tiempo transcurrido desde el inicio de la intervención hasta la aparición del evento.

No debemos olvidar que los pacientes reclutados en un ensayo son una muestra de los pacientes existentes o de los que veremos en el futuro y que nuestros resultados son solo estimaciones en esa muestra. Necesitamos información sobre el grado de reproducibilidad de esas estimaciones (precisión). Esta información viene generalmente suministrada por los intervalos de confianza. Para nuestra discusión baste comentar que nos dan una orientación sobre el nivel de confianza que podemos depositar en que la verdadera magnitud del efecto de la intervención se encuentre en los valores comprendidos entre los dos límites del intervalo. Por costumbre se utilizan niveles de confianza del 95%, asumiendo que es razonable aceptar una probabilidad de uno entre veinte (5%) de que dicho valor se escape de los límites del intervalo, pero este nivel puede perfectamente modificarse y adaptarse a necesidades o perspectivas particulares.

Tradicionalmente se han utilizado por los autores (y demandado por los editores y lectores) los valores p como indicadores de la significación estadística de los resultados. Existe una tendencia hacia una utilización creciente del estimador del efecto con su intervalo de confianza para informar de forma simultánea sobre la magnitud del efecto observado, su precisión y la existencia o no de significación estadística de los resultados. Como ejemplo ilustrativo presentamos los resultados del análisis del desenlace principal de varios ECA hipotéticos (tabla 5.2).

APLICABILIDAD

Hasta el momento se han revisado los principales aspectos metodológicos que pueden condicionar la validez *interna* de un ECA, es decir, aquellos aspectos que pueden llevar a cuestionar su calidad desde el punto de vista epistemológico. Pero la lectura crítica, como herramienta básica de la práctica

TABLA 5.2 ■ **Ejemplo de cuatro ensayos clínicos aleatorios hipotéticos**

Ensayo	Tratamiento (n)	Desenlaces	DR (IC 95%)	RR (IC 95%)
1	Experimental (100)	60	0,2 (0,06-0,34)	1,5 (1,12-2,00)
	Referencia (100)	40		
2	Experimental (30)	18	0,2 (–0,05-0,45)	1,5 (0,89-2,54)
	Referencia (30)	12		
3	Experimental (100)	55	0,05 (–0,09-0,19)	1,11 (0,84-1,46)
	Referencia (100)	50		
4	Experimental (2.000)	1.100	0,05 (0,02-0,08)	1,11 (1,04-1,18)
	Referencia (2.000)	1.000		

El desenlace de interés es favorable. Los ensayos 1 y 2 muestran cómo un efecto importante (grande) puede no ser detectado como estadísticamente significativo en un ensayo pequeño (ensayo 2). Los ensayos 3 y 4 muestran cómo un efecto moderado o pequeño puede alcanzar significación estadística si el ensayo tiene el suficiente tamaño (ensayo 4). Todos los casos muestran cómo los IC proporcionan simultáneamente información sobre la *precisión* de los resultados y sobre su (o la ausencia de) *significación estadística.* Obsérvese la inexistencia de valores *p* en la tabla.

DR, diferencia de riesgos; IC 95%, intervalo de confianza al 95%; RR, riesgo relativo.

Modificado de Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ. 1986;272:746-750.

basada en la evidencia, necesita ir un paso más allá para traspasar la frontera que separa el *pensar* del *hacer,* y para ello es imprescindible que el lector analice la aplicabilidad de los ECA a su actividad clínica real explorando la validez *externa* de sus hallazgos.

Entendemos por aplicabilidad la medida en que los efectos observados en los estudios publicados, probablemente, reflejen los resultados esperados cuando una intervención específica se aplique a la población de interés bajo condiciones de práctica real. Dicho de otra manera, la aplicabilidad debe intentar responder a tres preguntas: ¿pueden los resultados del ensayo ser aplicados a mi paciente?; o mejor ¿es mi paciente tan diferente de los pacientes del ECA que sus resultados no pueden serle aplicados?; ¿es la intervención factible en mi medio?

Un elemento esencial de la aplicabilidad es que las poblaciones de los ensayos son en general muy seleccionadas por razones metodológicas y regulatorias, de modo que son algo diferentes de los pacientes «cotidianos», presentando, por ejemplo, mucha menos comorbilidad que los pacientes habituales en la clínica. Adicionalmente, los entornos de investigación son más restrictivos en el manejo de los pacientes, más estrictos en las mediciones, y posiblemente más complacientes en las visitas y cuidados, etc. En resumen, hay diferencias entre los ensayos y la práctica real. Para rellenar ese hiato entre los ensayos y la aplicación práctica han surgido estrategias que tratan de dar sentido real (y no solo investigacional) a los resultados de los ECA: son los llamados estudios de mundo real (*Real World Studies* [RWS]). Este grupo de estudios (24,25) incluye diseños experimentales, como los ensayos pragmáticos o los ensayos basados en registros, y una serie de estudios observacionales basados en actividad o en registros de diferente tipo (generales o específicos) con arquitecturas de cohortes retrospectivas, estudios no aleatorizados, etc., y que constituyen un complemento interesante de la evidencia obtenida de los ECA. Abordaremos algunos de estos aspectos en el capítulo 10 aunque el análisis de estos estudios escapa a los límites de esta obra.

Un aspecto final de la aplicación a la práctica de los resultados de los ECA es la necesaria corrección y adaptación de los estimadores del efecto (positivos o negativos) a los riesgos basales de cada paciente, y la incorporación de los valores (colectivo e individuales) al proceso de decisión, aspecto que trataremos a propósito de las recomendaciones (v. capítulo 18).

Artículo

Vik I, Bollestad M, Grude N, Bærheim A, Damsgaard E, Neumark T, et al. Ibuprofen versus pivmecillinam for uncomplicated urinary tract infection in women—A double-blind, randomized non-inferiority trial. PLoS Med 2018;15(5):e1002569. Disponible en: https://doi.org/10.1371/journal.pmed.1002569.

Plantilla CASPe contestada para este artículo concreto

En el cuadro 5.1 se muestra la plantilla CASPe contestada para este artículo concreto.

Resolución del escenario

El estudio no muestra efectividad del ibuprofeno respecto del antibiótico, por el contrario, muestra un aumento de las infecciones del tracto urinario en el grupo de ibuprofeno. La ausencia de evidencia de efecto positivo en un solo ensayo no es evidencia de no efecto, pero existen datos que sugieren efectos negativos graves, especialmente pielonefritis. No obstante, los eventos observados para ese desenlace son pocos y por ello ese resultado es poco preciso y muy sensible a pequeños cambios en la incidencia en los grupos. Necesitaríamos una revisión sistemática para estar más seguros.

En cualquier caso, en este momento no prescribiríamos ibuprofeno a Manuela.

CUADRO 5.1 ■ Evaluación crítica del artículo propuesto (plantillas CASPe)

1. ¿Se orienta el ensayo a una pregunta claramente definida? *Una pregunta debe definirse en términos de la población de estudio, la intervención realizada, la comparación y los desenlaces considerados (positivos y negativos).* *Escribe los desenlaces.* *Puntúa su importancia según GRADE (No relevantes: 1-3; Importantes: 4-6, y Críticos para la decisión: 7-9).*	Sí ✓ No sé No **P** → pacientes mujeres no embarazadas de 18-60 años con síntomas de infección del tracto urinario (ITU) no complicada, es decir, disuria combinada con aumento de la frecuencia urinaria o urgencia urinaria o ambas, con o sin hematuria visible. SE EXCLUYÓ a pacientes con duración de síntomas mayor de 7 días, alérgicas a penicilinas o ibuprofeno, lactantes, pacientes con signos de ITU superior, infección vaginal, diabetes, enfermedad renal, etc. La detección de anormales en orina (tira reactiva para leucocitos, nitritos, proteínas y sangre) y el cultivo se hicieron a todas las pacientes, pero NO fueron criterio de inclusión ni de exclusión. **I** → Ibuprofeno 600 mg/8 h/3 días. **C** → Pivmecilinam 200 mg 3 veces al día × 3 días. **O** → El desenlace principal considerado fue la proporción de pacientes que se sintieron *curadas a los 4 días* tal y como se recogía en el cuaderno de la paciente (o en entrevista telefónica) (importancia *GRADE 6*). Los desenlaces secundarios fueron: duración de los síntomas *(GRADE 5);* puntuación de 0 a 6 según intensidad de disuria, urgencia urinaria y frecuencia urinaria (rango 0-18) *(GRADE 6);* proporción de pacientes con bacteriuria positiva en el segundo control *(GRADE 4);* proporción de pacientes con necesidad de consulta médica en las 4 semanas de seguimiento *(GRADE 4);* proporción de pacientes que recibieron antibióticos en las 4 semanas de seguimiento *(GRADE 5); proporción* de pacientes que desarrollaron ITU superior: pielonefritis *(GRADE 7);* proporción de pacientes con efectos adversos *(GRADE 6).* El **desenlace primario se evaluó como de no inferioridad** y el **resto como de superioridad.**
2. ¿Fue aleatoria la asignación a los tratamientos? *¿Se generó adecuadamente la secuencia?* *¿Se mantuvo oculta la secuencia de aleatorización?* *¿Son iguales en línea basal?*	Sí ✓ No sé No La secuencia de aleatorización se generó mediante una lista obtenida por ordenador creada por un estadístico independiente usando bloques de tamaño de 2, 4, 6 u 8, estratificados por país. La lista con los números de aleatorización correspondientes a cada centro se mantuvo centralizada y se conoció únicamente al final del estudio. Cada centro disponía de sobres opacos cerrados por si hubiera sido necesario descubrir el enmascaramiento. Cada grupo de 9 cápsulas de pivmecilinam o ibuprofeno se identificó con un número generado de acuerdo con la secuencia generada por el ordenador. Tras la inclusión, cada paciente recibía un kit numerado con las 9 dosis en el interior, de forma correlativa en cada centro. Los grupos (181 ibuprofeno y 178 pivmecilinam) estaban bastante bien balanceados (tabla 1), sin diferencias apreciables entre grupos.

3. ¿Se mantuvo la comparabilidad de los grupos a través del estudio? *Desviaciones de la intervención por problemas en la asignación.* *Desviaciones de la intervención por problemas en la adhesión al tratamiento.* *¿Cómo se analizó el estudio: ITT mITT APP, AT?*	Sí ✓	No sé	No
	Se aleatorizaron 383 pacientes, 194 al grupo ibuprofeno (IB) y 189 al grupo pivmecilinam (PIV). No hubo desviaciones en la intervención por problemas en la asignación. Los problemas de adherencia (< 80% de cumplimiento) fueron bajos, probablemente debido a la corta duración del tratamiento (3 días): 12 pacientes en el grupo IB y 7 en el grupo PIV. Se declaró análisis por intención de tratar (AIT o ITT) en las tablas, no en el texto, pero se analizaron únicamente 181 de 194 pacientes en el grupo IB y 178 de 189 en el grupo PIV. Podría considerarse un mITT *(modified intention to treat analysis)*. En realidad, y dado que el desenlace primario se analizó como de no inferioridad, lo razonable hubiera sido un análisis por protocolo (APP), lo que hubiera incluido 150 pacientes en el grupo IB y 154 en el grupo PIV (fig. 1). En realidad, así lo expresa (de modo confuso) en el pie de la tabla 2.		
4. ¿Son importantes las pérdidas ocurridas durante el estudio? *¿Difieren según el grupo?* *¿Las pérdidas podrían depender de su valor o resultado?* *¿Se hace análisis de sensibilidad?*	Sí	No sé	No ✓
	Se perdió a 44 de 194 pacientes en el grupo IB (22,7%): 13 de los que no se recuperó información tras la basal, 19 perdidos para el seguimiento y 12 que tuvieron baja adherencia. En el grupo PIV se perdieron 35 (18,5%): 11 de los que no se recuperó información tras la basal, 17 perdidos para el seguimiento y 7 que tuvieron baja adherencia. Puede considerarse que las pérdidas están balanceadas entre los grupos. Las diferencias de pérdidas son relevantes y mayores en el grupo IB, es posible que esos pacientes hayan buscado alternativa, y eso podría tener relación con un resultado negativo. No se hizo análisis de sensibilidad *(worst case, best case)*.		
5. ¿Fue adecuada la medición de los desenlaces? *Tipo de desenlace medido y método usado.* *Cegamiento del paciente, clínico, evaluador, estadístico.* *Si hay problema, ¿es diferencial entre los grupos?*	Sí	No sé ✓	No
	El desenlace primario medido fue la proporción de pacientes que se sintieron curados en el día 4 tal y como recogieron en el diario del paciente o se decidió tras la consulta telefónica. Los desenlaces secundarios incluyeron la duración de los síntomas y una puntuación de los síntomas del paciente reflejados en el diario según una escala *ad hoc* (mínimo 0, máximo 18). Otros desenlaces secundarios fueron proporción de pacientes con segundo cultivo positivo, proporción de pacientes con necesidad de consulta médica en las 4 semanas de seguimiento, proporción de pacientes que recibieron antibióticos durante este período. También se evaluaron desenlaces de seguridad: desarrollo de infección urinaria superior y efectos adversos. No queda muy claro cómo se generó el ciego. Parece (¿?) que una compañía farmacéutica «sobreencapsuló» el IB y el PIV, lo que parece querer decir que se introdujeron los comprimidos dentro de una cápsula igual para ambos. Se analizó que ambos preparados tuvieran el mismo aspecto, peso y sabor. Nos queda la duda de si hubiera podido averiguarse la intervención abriendo la cápsula, ya que hubiera quedado al descubierto la forma farmacéutica original de IB o PIV, que podría ser conocida de antemano por los pacientes o los investigadores. En todo caso, dado que la variable principal es mixta autorreportada + teléfono, la posible ruptura del ciego podría haber aumentado claramente el riesgo de sesgo. Aunque es hipotético, esa ruptura podría haber condicionado las pérdidas de la pregunta anterior.		

(Continúa)

6. ¿Se evitó la comunicación selectiva de resultados? *(Mirar el registro de ensayos)* *¿Hay reporte selectivo de desenlaces o reporte selectivo de análisis?*	Sí ✓ No sé No El ensayo aparece en clinicaltrials.gov con el n.º de protocolo NCT01849926 (https://clinicaltrials.gov/ct2/show/record/NCT01849926). Los desenlaces del punto anterior aparecen todos (excepto la carga de síntomas), pero además en el protocolo aparecen otros no recogidos en el ensayo: proporción de pacientes que tuvieron recaída en las 4 semanas de seguimiento y proporción de pacientes con cultivo positivo al cabo de 4 semanas.
7. ¿Cuál es el efecto del tratamiento para cada desenlace? *¿Qué desenlaces se han medido?* *Detalla los positivos y los negativos.*	**Desenlace principal:** En el grupo IB 70 pacientes (38,7%) se sintieron curados el día 4, frente a 131 (73,6%) en el grupo PIV (RAR 35,3%; IC 95% 25,7-44,9; NNT = 3; IC 95% 3-4). Este valor está fuera del margen de no inferioridad. Por tanto, IB no demostró la no inferioridad frente a PIV. **Desenlaces secundarios:** Duración media de los síntomas tras la aleatorización → 6 días IB vs. 3 días PIV. Pacientes sin síntomas el día 7 → 114 (63%) en IB vs. 162 (91%) en PIV (RAR 28; IC 95% 20-36). Pacientes sin síntomas el día 14 → 141 (78%) en IB vs. 167 (94%) en PIV (RAR 16; IC 95% 9-23). Cultivo de orina positivo a los 14 días → 43 (28%) en IB vs. 16 (10%) en PIV (RAR 16; IC 95% 7-26). Tratamiento con antibióticos el día 14 → 73 (41%) en IB vs. 14 (8%) en PIV (RAR 32; IC 95% 24-40). Tratamiento con antibióticos el día 28 → 83 (46%) en IB vs. 18 (10%) en PIV (RAR 36; IC 95% 27-44). Pacientes con ITU febril → 5(3%) en IB vs. 0 (0%) en PIV (RAR 3; IC 95% 0,1-6). Pacientes con pielonefritis → 7(4%) en IB vs. 0 (0) en PIV (RAR 4; IC 95% 1-8). Efectos adversos graves → 6 (3%) en IB vs. 1 (1%) en PIV (RAR 3; IC 95% 6 a –0,1). En resumen: el IB es claramente inferior para control de síntomas y claramente superior para producir ITU + fiebre o pielonefritis (que son indeseables).
8. ¿Cuál es la precisión de los estimadores del efecto? *¿Cuáles son sus intervalos de confianza?*	Ver apartado anterior. Dado que la diferencia de efecto es en general grande y la muestra razonablemente amplia, los IC son estrechos, lo que concede fiabilidad a los resultados.

9. ¿Pueden aplicarse estos resultados en tu medio o población local? *¿Crees que los pacientes incluidos en el ensayo son demasiado distintos a tus pacientes?* *¿Hay algún otro ensayo parecido a este?* *¿Es consistente con este?*	Sí	No sé ✓	No
	Nuestra paciente del escenario hubiera cumplido criterios de inclusión, pero no hubiera podido ser incluida en este ensayo por su alergia a betalactámicos. Tal vez pudiera haber sido incluida en alguno de los ensayos similares llevados a cabo con quinolonas y que tuvieron resultados parecidos al que nos ocupa (v. más adelante). En la Introducción, los autores mencionan otros artículos similares: un pequeño estudio piloto en Alemania (ref. 13 del artículo) en el que IB fue no inferior a ciprofloxacino; otros dos más amplios en Suiza y Alemania (ref. 14 y 15 del artículo) mostraron que el tratamiento antibiótico (norfloxacino y fosfomicina respectivamente) fue superior a los antiinflamatorios no esteroideos (diclofenaco e IB, respectivamente). Por tanto, hay consistencia con los hallazgos de este estudio.		
10. ¿Se han tenido en cuenta todos los resultados y su importancia clínica? *Utilidades y disutilidades de cada desenlace.* *Balance de efectos positivos/negativos.* *Preferencias del paciente, costes, etc.*	Sí	No sé	No ✓
	Sí se tuvieron en cuenta las utilidades (beneficio clínico) y las disutilidades (efectos adversos graves), por otra parte nada frecuentes. El IB es claramente inferior para control de síntomas y claramente superior para producir ITU + fiebre o pielonefritis (que son indeseables). Pero además hay diferencia entre la utilidad de quitar síntomas menores, que es poca, y la disutilidad de ITU o pielonefritis, que es mucha, y más si se valora a largo plazo. Sin embargo, pensamos que la respuesta ha de ser negativa en este punto porque, aunque se contemplaron aspectos clínicos, no se tuvieron en cuenta los costes ni las preferencias de los pacientes.		
11. ¿Los beneficios que se espera obtener justifican los riesgos y los costes? *Es improbable que pueda deducirse solo de un ensayo, pero ¿qué piensas tú al respecto?*	Sí	No sé	No ✓
	Dado que IB no ha mostrado no inferioridad con respecto a PIV (lo que puede entenderse a efectos prácticos como que ha demostrado inferioridad), es evidente que los beneficios con su uso no justifican en absoluto los costes. **CLARAMENTE NO.** Discrepamos del comentario de los autores en la conclusión del estudio, que insisten en las ventajas del IB con base en argumentaciones colaterales.		

Bibliografía

1. Frieden TR. Evidence for Health Decision Making – Beyond Randomized, Controlled Trials. N Engl J Med 2017;377(5):465-75.
2. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007;334(7589):349-51.
3. Higgins JPT, Savovic J, Page MJ, Elbers RG, Sterne AC. Assessing risk of bias in a randomized trial. En: Higgins JPT, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editores. Cochrane Handbook for Systematic reviews of interventions. 2.ª ed. Hoboken: Wiley & Sons; 2019. p. 205-28.
4. Cabello JB, Abraira V, Gómez J. Ensayos clínicos para un solo paciente. Justificación, metodología y aspectos bioéticos. Med Clin Barc 1997;109:592-602.
5. Montori VM, Devereaux PJ, Adhikari NKJ, Burns KEA, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA 2005;294(17):2203-9.

6. Cannistra SA. The ethics of early stopping rules: who is protecting whom? J Clin Oncol Off J Am Soc Clin Oncol 2004;22(9):1542-5.
7. Montori VM, Permanyer-Miralda G, Ferreira-González I, Busse JW, Pacheco-Huergo V, Bryant D, et al. Validity of composite end points in clinical trials. BMJ 2005;330(7491):594-6.
8. McPherson GC, Campbell MK, Elbourne R. Use of randomization in clinical trials. Trials 2012;13:198.
9. Altman DG, Bland JM. Treatment allocation by minimisation. BMJ 2005;330(7495):843.
10. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet Lond Engl 2002;359(9306):614-8.
11. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273(5):408-12.
12. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336(7644):601-5.
13. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLOS ONE 2016;11(7):e0159267.
14. Moustgaard H, Clayton GL, Jones HE, Boutron I, Jørgensen L, Laursen DRT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ 2020;368:l6802.
15. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. BMJ 2006;332(7547):969-71.
16. Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol 2013;13:92.
17. Naci H, Davis C, Savović J, Higgins JPT, Sterne JAC, Gyawali B, et al. Design characteristics, risk of bias, and reporting of randomised controlled trials supporting approvals of cancer drugs by European Medicines Agency, 2014-16: cross sectional analysis. BMJ 2019;366:l5221.
18. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. PLoS ONE 2008;3(8):e3081.
19. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. Clin Trials J Soc Clin Trials 2012;(1):48-55.
20. Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. N Engl J Med 2017;377(14):1391-8.
21. Hernán MA, Scharfstein D. Cautions as Regulators Move to End Exclusive Reliance on Intention to Treat. Ann Intern Med 2018;168(7):515.
22. Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
23. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.
24. Fanaroff AC, Steffel J, Alexander JH, Lip GYH, Califf RM, Lopes RD. Stroke prevention in atrial fibrillation: re-defining «real-world data» within the broader data universe. Eur Heart J 2018;39(32):2932-41.
25. Dal-Ré R, Janiaud P, Ioannidis JPA. Real-world evidence: How pragmatic are randomized controlled trials labeled as pragmatic? BMC Med 2018;16(1). Disponible en: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-018-1038-2.

Cómo citar este capítulo:

Cabello JB, López E, Pijoan JI. Lectura crítica de estudios de tratamiento. Ensayos clínicos aleatorios. En: Cabello Juan B, editor. Lectura crítica de la evidencia clínica, 2.ª ed. Barcelona: Elsevier; 2022. p. 36-56.