MEDUNAB

REVISIÓN DE TEMA

Elementos para la valoración y el uso práctico de los ensayos clínicos
Parte I: Estructura básica y terminología usual

Juan Carlos Villar Centeno*

RESUMEN

Los ensayos clínicos (EC) constituyen la base para la creación de recomendaciones terapéuticas basadas en la evidencia. Por lo tanto, es deseable que la comunidad académica local aumente su preparación y habilidades en el análisis crítico de EC para juzgar su validez, valorar el tamaño de sus resultados y determinar la aplicabilidad a su propia práctica. Este artículo (la primera de dos partes) presenta los fundamentos conceptuales y las características más importantes de los EC, a la vez que presenta algunos términos de uso corriente en los informes de este tipo de estudios. A lo largo de este artículo se discuten: a) Los orígenes y consecuencias del error alfa y beta en los EC; b) El principio de la aleatorización y la evaluación del balance de los grupos en la línea de base; c) Las maneras de establecer comparaciones entre los grupos de tratamiento como base de las inferencias sobre el efecto de las intervenciones.

Palabras clave: Ensayos clínicos, metodología, aleatorización**.**

* Profesor asociado, Facultad de Medicina, Universidad Autónoma de Bucaramanga, Bucaramanga, Colombia. Research Fellow, Population Health Institute; Registered Student, Clinical Health Sciences (Health Research Methodology) Graduate Program, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canadá.

Corrrespondencia. Dr. Villar, Hamilton General Hospital McMaster Clinic – Population Health; 238 Barton St. East, Hamilton ON, L8L 2X2, Canadá. e-mail: juan@ccc.mcmaster.ca

Introducción

La publicación de un ensayo clínico (EC)[1] en el pasado número de MEDUNAB dedicado al dolor constituye una importante novedad que merece el reconocimiento de la comunidad académica local a los autores. Este importante avance también sirve de estímulo para investigadores, editores y usuarios de la investigación en nuestra comunidad. Para los lectores, al igual que con otros estudios, un EC plantea el reto de saber evaluar críticamente la validez que su diseño y conducción demuestran, apreciar la magnitud relativa de sus resultados y determinar la aplicabilidad en su práctica.[2] Sin embargo, con los EC esta responsabilidad podría ser un poco mayor que la que podría tenerse frente a estudios diagnósticos o pronósticos, debido a su mayor potencial de modificar las intervenciones que el equipo de salud práctica.[3] Por ejemplo, en este EC los autores evaluaron la utilidad de modificar una técnica anestésica actualmente en uso y concluyeron que su estudio "demuestra la utilidad de adicionar opioides a la mezcla utilizada en anestesia peribulbar en cirugía de catarata", invitando a una modificación de la práctica clínica en esa dirección. Así, la publicación de este EC en MEDUNAB brinda una oportunidad al medio para hacer una discusión sobre los aspectos vertebrales de los EC y presentar algunos de los términos que usualmente son empleados en sus informes y los elementos de importancia en la evaluación de la validez de sus resultados.

El presente artículo parte del estudio en mención como punto de referencia, pero a través de él pretende introducir una visión aplicable a la generalidad de los EC con el objetivo de aumentar los elementos de juicio que permitan a los lectores acceder, interpretar y usar con mayor eficacia estos estudios, es decir, hacia la construcción de una práctica basada en la evidencia. Dividiremos este artículo en dos partes: en la primera se presentan los aspectos fundamentales de la estructura de los EC y se definen los términos más importantes (subrayados a lo largo del texto y presentados junto al término usado en las publicaciones en *inglés*), de modo que el usuario pueda ubicarse adecuadamente en dicho escenario. En la segunda se darán los elementos para evaluar la calidad del diseño y el informe de los EC, de modo que el usuario pueda juzgar su validez y con ello decidir la aplicabilidad de estos en su propia práctica.

El ensayo clínico como juicio de responsabilidad con la práctica

Un modelo imperfecto con altas implicaciones en la práctica. Aunque la literatura hispana ha elegido la acepción de la palabra inglesa 'trial' como 'ensayo', me gustaría usar para efectos de esta discusión la acepción 'juicio',[4] con el fin de hacer una analogía que permita enfatizar algunos conceptos. Estos juicios se realizan para decidir cual tratamiento produce mejores resultados en una población determinada. El jurado en este caso (cada lector) decide con base en los argumentos presentados por los abogados (los investigadores), si hay mérito suficiente en ellos como para sugerir al juez que produzca una nueva sentencia que modifica una situación (la práctica clínica). Los argumentos que los abogados presentan a los miembros del jurado en su intervención (el informe del estudio) son, sin embargo, el producto de su interpretación de esa evidencia. Por tanto, en razón a su naturaleza humana, unos y otros están expuestos (consciente o inconscientemente) a sesgos (bias) de diverso origen, que hacen que la generación de la información, su transmisión en el informe o su interpretación no corresponda a los hechos. Por supuesto, a través del tiempo el sistema de justicia se ha perfeccionado generando las reglas del juego (el 'debido proceso') buscando reducir al máximo estos sesgos en el entendido de la imposibilidad de anularlos. Así, los investigadores y los usuarios de la investigación comparten una responsabilidad ética y académica, de modo que se garantice, mediante el seguimiento riguroso de aquellas normas que aseguran un juicio justo, que la mayor cantidad posible de sesgos sean evitados o minimizados en favor de los pacientes y los sistemas de salud.

Administrar justicia o presenciar juicios perfectos no es sin embargo tarea fácil. Por eso no es posible garantizar que en todos los casos las sentencia emanada de un solo EC se convierta en dogma irrevocable. Raramente las recomendaciones terapéuticas surgen de un solo EC, sino que usualmente se busca que los resultados sean confirmados por otros EC que exploran la misma pregunta de investigación.[5] Si la evidencia de un EC es generada, interpretada y juzgada certeramente, el resultado será la construcción de conocimiento y la irradiación de ésta en la práctica de medicina basada en la evidencia. Si por el contrario durante estos procesos se presentan sesgos, entonces las consecuencias serán adversas desde diferentes puntos

de vista, según la situación. Por ejemplo, cuando el producto final de un EC no representa la realidad (pese a la mejor intención de los investigadores) puede sugerirse erróneamente que cierta intervención reduce el número de situaciones indeseables para la salud de los participantes. Con esto se puede introducir en la práctica clínica una intervención que no funciona, aumenta el costo de atención o las expectativas de pacientes y la sociedad e incluso retrasa la generación de otras alternativas o hipótesis con potencial utilidad.

Los ensayos clínicos probando la administración de magnesio para el tratamiento del infarto del miocardio pueden servir como ejemplo. Un meta-análisis que combinaba los hallazgos de los primeros siete EC (con un total de 1,301 participantes) concluyó que este constituía un coadyuvante útil para esos pacientes.[6] Tiempo después se produjeron nuevos EC (con más de 50.000 participantes) evaluando la misma terapia,[7] en donde se evidenció un efecto neutro de esta intervención. La discordancia en los resultados de estos EC han generado intenso debate y especulación académica, al punto que todavía algunos consideran este tópico no resuelto aún,[8] después de más de una gran inversión de tiempo y recursos.

La situación contraria también puede darse: los investigadores pueden declarar la no existencia de diferencias, cuando en realidad ellas existen. Esta conduciría a que una intervención benéfica deje de implementarse oportunamente durante un tiempo indeterminado, o mucho peor aún, que una intervención con un potencial dañino continúe siendo utilizada. Ejemplo de esto lo constituyen los EC que no encontraron diferencia en la incidencia de hemorragias cerebrales al comparar la administración de trombolíticos en infusión o en bolo, concluyendo que estas eran intervenciones equivalentes respecto al efecto terapéutico y a los efectos secundarios. En vista de la mayor agilidad y facilidad para su administración, la administración en bolo debería ser la elección. Sin embargo, un meta-análisis posterior sumando los datos contenidos en siete estudios que abordaban esta pregunta demostró un exceso comparativo de hemorragias intracerebrales en los pacientes que habían recibido trombolíticos en bolo.[9] Así, las implicaciones éticas, médicas y económicas de estos juicios que lleven a decisiones erróneas pueden causar enorme impacto para individuos o colectivos, dada la influencia más directa de los EC sobre la práctica cotidiana. Por ello, el usuario de un EC deberá ser cuidadoso en considerar la aplicación directa de los resultados de un EC en su práctica sin examinar cuidadosamente la presencia de circunstancias que amenacen su validez y sin ubicar estos resultados en el contexto del conocimiento sobre el tópico y los resultados de otros EC que aborden preguntas de investigación similares.

El azar como fuente de error en los ensayo clínicos. Además de la introducción de sesgos a lo largo del proceso, el azar mismo puede jugar un papel en la desviación de los resultados de la realidad.[10] Usando los dos escenarios anteriores, el juego de las probabilidades puede llevar a dos situaciones potenciales, ambas conducentes a una interpretación errónea. A casos como el del uso del magnesio, la jerga estadística los denomina error alfa (o tipo I), la situación en la que se identifican diferencias en donde en realidad no existen, en este caso, por el simple azar. Esto podría ocurrir si el marco muestral (*sampling frame*), el (los) centro (s) o servicio (s) de donde se toma la muestra, o la muestra misma de participantes no representa todo el espectro de la población objeto (*target population*), los pacientes sobre los que se supone se aplican los resultados. Puede esto darse en un EC bien por su tamaño, por un período relativamente corto de recolección de los datos (si la enfermedad en cuestión tiene algún patrón temporal) o por la inclusión de un número de centros relativamente insuficiente para cubrir ese espectro (si existen por ejemplo diferencias geográficas en el espectro de la enfermedad).

En el caso de la situación de los trombolíticos, la situación inversa, el término es error beta (o tipo II), cuando no se detecta una diferencia que en realidad existe. El error beta es fuertemente dependiente del tamaño de la muestra de participantes. Por tanto, la mejor manera de evitarlo será la incorporación de un número suficiente de participantes para la pregunta de investigación en cuestión. El tamaño de la muestra es usualmente definido con antelación, con base en lo que los investigadores consideren como una diferencia clínicamente importante que se desee demostrar o, en este caso, negar de manera confiable. De ahí que un EC de gran tamaño (algunos han categorizado los EC en pequeños, medianos y grandes si tienen menos de 500, entre 500 y 2.000 a más de 2.000 participantes respectivamente), cuyo período de recolección de datos sea relativamente largo y en el que participe un espectro amplio y numeroso de centros reclutando participantes, tendrá una menor probabilidad de que el juego del azar o el tamaño insuficiente de la muestra les hagan incurrir en estas dos situaciones.[11] Se ha

demostrado también que los EC con estas características tienen una mayor probabilidad de ser informados en la literatura, contienen una menor cantidad de sesgos en su diseño y conducción, y su informe es de mejor calidad.[12]

De este modo, el investigador debe aceptar el reto, como generador y presentador de los argumentos jurídicos, de generar el mejor ensayo clínico posible en términos de minimizar los sesgos y el papel del azar en sus resultados, extremando su eficiencia y posibilidades logísticas. El lector, por su parte, como miembro del jurado, tiene la responsabilidad de visualizar y calificar todo aquello que potencialmente lo conduzca a un juicio equívoco, contado con el máximo posible de elementos de juicio para analizar críticamente la información que le es presentada para identificar, calificar y dar peso a estos sesgos en su decisión final.

Características fundamentales de un ensayo clínico

En el entendido de que todas las intervenciones evaluadas merecen un juicio justo, la manera como se conduzca este proceso será entonces crítica para asegurar el objetivo final de hacer justicia (practicar la medicina basada en la evidencia). Dos principios relativamente simples constituyen la columna vertebral del proceso de los EC: la asignación aleatoria de la intervención en estudio y la comparación de sus efectos en el grupo experimental frente a un grupo control. Habida cuenta de tales principios, llamamos 'ensayo clínico' (*clinical trial*) a lo que suponemos es en realidad un "ensayo clínico aleatorizado controlado" (*randomized controlled trial*). Con esas reglas del juego, nuestra decisión final la deducimos de la existencia o no de diferencias entre los efectos observados en ambos grupos sometidos a tratamiento. ¿Por qué entonces principios aparentemente simples de observar pueden tornarse tan complejos en la práctica?

Asignación aleatoria de las intervenciones. La comparación de los efectos de dos tratamientos podría enfrentar el problema de que los grupos a ser tratados fueran diferentes en muchos aspectos entre si, lo que en la expectativa de hacer una comparación final podría generar diferencias no atribuibles a la intervención bajo experimentación, sino a factores ajenos a ésta.

Por ejemplo, si estamos comparando el efecto de un nuevo analgésico opiaceo entre dos grupos de personas que difieren sustancialmente en la frecuencia de uso previo de otros analgésicos opiaceos, no podríamos aseverar fácilmente que una diferencia en la dosis analgésica requerida fuera atribuible al efecto de la nueva medicación. Esto es, una conclusión de mayor efectividad analgésica (en términos de la dosis necesaria para controlar el dolor) en el grupo que al inicio del estudio tenía una menor prevalencia de uso de opiaceos podría estar claramente sesgada. En este ejemplo el uso previo de opiaceos constituyó un factor que, por no estar presente por igual en ambos grupos, pudo haber confundido esa comparación. Por tanto, se requiere que los grupos que serán expuestos a tratamiento mantengan un balance respecto a todas estas variables de confusión que potencialmente pueden afectar los resultados.

Por definición, una variable de confusión está presente en la población expuesta al factor en estudio (en este caso el tratamiento) y tiene a la vez la posibilidad de modificar la respuesta que se está evaluando. La presencia desigual (carencia de balance) de estas variables en los grupos a ser comparados se denomina sesgo de selección. Un investigador debe por lo tanto buscar (y el lector confirmar) el balance de los grupos antes de atribuir un efecto al tratamiento que se está probando. Sin embargo, estas variables pueden llegar a ser muy numerosas o difíciles de controlar para el investigador. Aún en el caso que este pudiera hacerlo, podría no visualizar o incluso no conocer todas estas variables. ¿Cómo podríamos entonces garantizar un balance de todos los factores de confusión entre los grupos a ser comparados? Al distribuir en proporciones aproximadamente iguales la prevalencia de estas variables entre los participantes asignados a los tratamientos, la aleatorización (*randomization*) constituye la herramienta más simple y poderosa para lograr el mayor balance posible de los factores de confusión (conocidos y desconocidos) presentes en los grupos de estudio y con ello disminuir los sesgos de selección.

La aleatorización, la ausencia de patrones predecibles en una secuencia, necesita un método confiable que asegure este objetivo. Entre los métodos utilizados por los investigadores debe destacarse que procedimientos como la selección del primero de cada tres pacientes que asisten a una consulta, los sujetos que son hospitalizados los lunes, miércoles y viernes, o la selección de aquellos animales de laboratorio que toman el camino de la derecha constituyen ejemplos de seudoaleatorización, pues con estos no es posible excluir patrones no siempre aparentes que determinan esa distribución. En contraste, métodos

aún simples como el lanzamiento de una moneda (si se desea asignar tratamiento a dos grupos), de un dado, el uso de una tabla de números aleatorios (como lo hicieron Rangel y cols[1]), o el uso de calculadoras o computadores para la generación de secuencias aleatorias constituyen métodos aceptables hacia ese objetivo. Aunque la mayoría de los EC publicados son presentados como aleatorizados, la descripción de los métodos de aleatorización se presenta en 20-40% de los EC[2;3]. Esta proporción aplica también a las publicaciones más sobresalientes, en donde en una muestra de 80 EC publicados en cuatro de las revistas de mayor influencia no se encontró evidencia de aleatorización en un 30% de los estudios[4]. En nuestro ejemplo, los autores presentaron su estudio como aleatorizado y mencionaron el método en el informe (ver sección "materiales y métodos" de la publicación).

Una asignación verdaderamente aleatoria de tratamientos a grupos suficientemente grandes de participantes haría innecesaria una prueba estadística formal para demostrar la no existencia de diferencias entre los grupos en la línea de base. Sin embargo, la mayoría de informes de los EC presentan la distribución de ciertas variables de interés para cada grupo acompañadas de una prueba estadística para demostrar la ausencia de diferencias entre esos grupos. Dicha información está usualmente presentada en la primera tabla de los informes (típicamente titulada "*baseline characteristics*") y contiene aquellas variables que a juicio de los autores tendrían el potencial de convertirse en variables de confusión para los resultados, en caso de no tener balance apropiado al inicio del estudio. Se desea con esto confirmar que se ha logrado el objetivo de alcanzar una distribución balanceada de las variables de interés entre los grupos. Sin embargo, más allá de lo estipulado por los autores en su discusión, el usuario de un EC debería examinar la presencia de ese balance en el contexto del diseño del estudio en cuestión. De hecho, el estudio citado arriba en donde se estudiaron las características de la aleatorización de los EC[4] encontró que en los informes analizados, la estadística usada (para la demostración de no diferencias) fue inapropiada en un 41% de la muestra examinada.

En nuestro ejemplo, Rangel y cols[1] presentan los aspectos demográficos y las características clínicas de base de los participantes (tablas 1 y 2 del informe), acompañados de la prueba estadística para la comparación entre los grupos de tratamiento. Se describe allí que no se encuentran diferencias "estadísticamente significativas", lo que en muchos estudios, incluido el de Rangel y cols,[1] se define como un valor de $p > 0.05$ para la prueba de la hipótesis de no diferencias entre los grupos. Este valor significa que existe un 5% o más de probabilidad de que los grupos que son comparados sean iguales. En esas tablas, en donde se describe la prevalencia de patología asociada en los diferentes grupos asignados en la línea de base, se informa un valor de p para varias de las comparaciones (exactamente en 5 de 12) que es mayor del 5%, pero menor del 50%. Habiendo usado un método aceptable de aleatorización, ¿no se esperaría entonces que la distribución de esas variables entre los grupos fuera aún más similar y, por lo tanto, el valor de p un poco más alto?

Una explicación radica, desde luego, en el juego del azar en contra de una distribución completamente balanceada. La situación de falta de balance se minimiza progresivamente, a medida que el número de pacientes en quienes es 'sorteada' la asignación de tratamientos crece. En la tabla 1 se muestra cómo los intervalos de confianza del 95% (el margen de fluctuación alrededor del valor estimado que podría resultar el 95% de las veces bajo esas condiciones), para una proporción teórica del 50% van siendo más 'estrechos' mientras aumenta el tamaño de la muestra del estudio.[16] Esta situación equivale a lanzar una moneda al aire, en donde a medida que aumentamos el número de lanzamientos, el número de veces que obtenemos una 'cara' se acerca al 50% exactamente. Dicho de otra forma, a medida que se aumenta el número de 'sorteos' la influencia del azar en desviar el número de eventos de la proporción esperada es cada vez menor. En el caso que nos ocupa, la probabilidad de sesgo de selección es inversamente proporcional al número de sujetos en los que se asigna aleatoriamente el tratamiento.

Además de la influencia crítica del tamaño de la muestra, el número de grupos comparado también influye en el balance de las variables de confusión entre grupos en la línea de base, lo que se refleja en el valor de p. Rangel y cols[1] diseñaron su estudio para asignar tres tratamientos al azar, lo que daría una probabilidad teórica de asignación del 33.3% de individuos a cada grupo. Lo que en la práctica sucedió fue que de los 104 sujetos a los que se asignó una de las tres soluciones anestésicas, 34 (32.7%) fueron asignados al grupo A, 32 (30.8%) al grupo B y 36 (36.5%) al grupo C, lo que demuestra ese patrón esperado de distribución de la muestra aleatorizada.

La mayoría de EC grandes asignan su intervención a dos grupos. Sin embargo, muchos EC medianos y pequeños utilizan un número mayor de grupos de

Tabla 1. Cambios en los intervalos de confianza del 95% (IC 95%) para una proporción esperada de eventos del 50% (por ejemplo, la frecuencia con la que se obtiene una 'cara' al lanzar una moneda) a medida que el tamaño de la muestra (o de lanzamientos) aumenta.

TAMAÑO DE LA MUESTRA	PROPORCIÓN ESPERADA (%)	IC 95%	
		INFERIOR (%)	SUPERIOR (%)
10	50	27.1	72.8
20	50	33.8	66.2
40	50	38.6	61.4
80	50	42	58
160	50	44.4	55.6
320	50	46.1	53.9
640	50	47.2	52.7
1280	50	48	52
2560	50	48.6	51.3
5120	50	49	51

estudio. A medida que el número de grupos comparados aumenta, la probabilidad de demostrar diferencias entre ellos en la línea de base es menor. La tabla 2 muestra la evolución de los valores de p a medida que el número de grupos aumenta para un grado de disparidad entre grupos que se mantiene estable. El aumento de los grados de libertad es deducido del número de grupos comparados (en el caso de una evaluación de dos tratamientos estos serían igual al número de grupos comparados – 1). El valor de chi cuadrado es el resultado del grado de disparidad observado en la proporción de sujetos en cada grupo. Para el mismo chi cuadrado, a medida que el número de grados de libertad aumenta, el valor de p para probar la homogeneidad entre los grupos aumenta. La situación se repite en el ejemplo para dos valores diferentes de chi cuadrado. Así, la inclusión de un número mayor de grupos a ser comparados en un EC

Tabla 2. Valores de p para la hipótesis de no diferencia en la prevalencia de cierta variable cuando se compara un número creciente de grupos (representado en los grados de libertad). Se muestran dos escenarios en los que el grado de disparidad entre los grupos (representado en el valor de Chi cuadrado) se mantiene constante (2.5 o 5).

GRADOS DE LIBERTAD	CHI CUADRADO	VALOR DE p	CHI CUADRADO	VALOR DE p
1	2.5	0.1138	5	0.0253
2	2.5	0.2865	5	0.0821
3	2.5	0.4753	5	0.1718
4	2.5	0.6446	5	0.2873
5	2.5	0.7765	5	0.4159

daría una menor probabilidad de encontrar diferencias (falta de balance) entre los grupos en la línea de base. En nuestro ejemplo, el hecho de que algunos de los valores de p (bajo el título "significancia" en las tablas del artículo) sean más cercanos a 0 que a 1 es el resultado de un tamaño de muestra en donde los intervalos de confianza alrededor de la proporción esperada son aún amplios (lo que haría disminuir el valor de p), pero también de tener tres grupos a ser comparados (lo que haría aumentar el valor de p).

En conclusión, más allá de dar un 'visto bueno' a la ausencia de diferencias 'estadísticamente significativas', el lector debe interpretar esos valores a la luz del peso que un eventual desbalance de una o más variables de confusión entre los grupos a ser tratados pueda tener para las inferencias del estudio, de acuerdo con el método de aleatorización empleado, el tamaño de su muestra y el diseño en general del estudio. La observación del simple enunciado de que la asignación de tratamiento se haga al azar, aún usando un método aceptable de aleatorización (y haciéndolo evidente en el informe), no puede garantizar por si mismo al lector la carencia de sesgos de selección, o de errores alfa o beta en las conclusiones presentadas por los autores. Todas las inferencias posteriores que un usuario haga sobre un EC deben partir de ese marco inicial.

Comparación del efecto entre intervenciones. Una vez hecha la asignación de los tratamientos, la segunda característica de los EC es la comparación de los efectos de la intervención frente a un grupo control. Esta comparación se realiza evaluando la presencia o no de diferencias en los desenlaces *(outcomes o endpoints)* que el estudio ha presupuestado medir al final de un período de observación *(follow-up period)* requerido para la verificación de los cambios producidos por la intervención. Estos desenlaces pueden ser tan variados como cambios anatómicos, fisiológicos, clínicos, incidencia de mortalidad, número de hospitalizaciones o cirugías, aparición o desaparición de conductas, cambios en el uso de procedimientos o medicamentos, costos, etc. de acuerdo con el contexto del estudio.

Los desenlaces se caracterizan por ser condiciones medibles que se supone la intervención en estudio puede modificar. Al final del período de seguimiento previsto se compara la cantidad de desenlaces ocurridos en los grupos tratados con las intervenciones. En el supuesto que los sesgos de selección hayan sido controlados (en realidad, minimizados) previamente, las diferencias que se puedan observar serían atribuibles a la "única" circunstancia que hace los grupos diferentes entre si: el efecto de la intervención en estudio. Dado que el proceso en cuestión no fue simplemente observado por los investigadores, sino modificado por estos con la introducción de un factor experimental (el tratamiento), los EC son estudios experimentales, en este caso experimentos clínicos.

Muchas veces en los EC se presupuesta el registro de diferentes categorías de desenlaces, de acuerdo con la jerarquía que para los investigadores tiene la resolución de las preguntas de investigación. Usualmente estos son llamados desenlaces primarios (*primary outcomes*) y desenlaces secundarios (*secondary outcomes*) y corresponden a las preguntas de investigación de primera o segunda importancia para los autores. En nuestro ejemplo, los desenlaces que el estudio midió fueron el tiempo de latencia y el tiempo de analgesia postoperatoria (ver definiciones en el informe), al igual que la incidencia de complicaciones intra y postoperatorias, los tiempos quirúrgicos y la necesidad de aplicación de bolos adicionales de anestesia peribulbar o de uso de analgésicos de rescate. Aunque Rangel y cols[1] no mencionan específicamente en su informe cuáles de esos desenlaces son para ellos 'primarios' y cuales 'secundarios', en el resumen se establece claramente que la evaluación del tiempo de latencia y el control postoperatorio del dolor son los objetivos (se asume, primarios) del estudio. Adicionalmente, los autores basaron su estimado del tamaño de la muestra en una diferencia hipotética en el tiempo de latencia, lo que permite suponer que entre los diferentes desenlaces medidos en el estudio, éste sería el más importante.

La comparación de esos desenlaces es la base para establecer la presencia de diferencias entre los grupos, y éstas, a su vez, la base para extractar inferencias respecto al efecto de la intervención. Las comparaciones son realizadas entre un grupo de tratamiento activo y un grupo control, que actuaría como sustraendo. Este último puede ser un grupo sometido a una intervención a la que se considera inerte biológicamente (placebo) o tratado con los patrones usuales de práctica que se comparan frente al efecto de la nueva intervención que se añade a esos patrones preexistentes. La selección de uno u otro grupo de comparación es el resultado del contexto clínico y de las consideraciones éticas alrededor de este. Por ejemplo, la administración de placebo como único tratamiento sería éticamente inaceptable si se conoce que no ofrecer un tratamiento al grupo de comparación conduce a desenlaces adversos. Otras veces se añade el tratamiento

experimental o un placebo al tratamiento que se considera convencional en una situación en la que los pacientes deben usar varios medicamentos simultáneamente.

Nuestro ejemplo compara un tratamiento convencional (la mezcla "A") frente a la adición de cualquiera de dos nuevos ingredientes a la mezcla convencional (morfina en la mezcla "B" y fentanyl en la mezcla "C") para generar tres intervenciones a ser comparadas. Aquí resulta obvia la inconveniencia ética de usar una mezcla que no tuviese efecto analgésico como grupo control. Dado que se administró una de las diferentes soluciones anestésicas a los participantes asignados a cada grupo en más o menos el mismo momento del proceso quirúrgico, se trata de una comparación de intervenciones realizadas paralelamente. Este tipo de diseño se llama EC paralelo. Rangel y cols[1] condujeron un EC con diseño paralelo con tres grupos de estudio.

Una situación alternativa de comparación se plantea en el diseño cruzado ("*cross-over design*"). En este tipo de EC, cada grupo es sometido primero a una intervención y luego de un período de lavado ("*washout period*") es administrada a una segunda intervención asignada al participante. En este caso, la aleatorización está dirigida a indicar una secuencia de intervenciones (v.g. AB, o BA,) más que una intervención específica (A o B) como en el diseño paralelo. Aquí, cada participante es parte del grupo activo y del grupo control en dos momentos diferentes del estudio. En otro tipo de diseño puede darse la comparación de los efectos de más de un tipo de intervención simultáneamente; por ejemplo, comparar los efectos de A versus B, pero también comparar C versus D en el mismo EC. Cada participante es aleatorizado a un 'brazo' de cada una de las intervenciones simultáneamente (A+C, A+D, B+C, etc). Este diseño se llama diseño factorial pues en este EC se está introduciendo más de un factor en el experimento (cada uno de los pares de tratamiento activo-control constituyen un factor). Si se evalúan dos factores, cada uno con dos 'brazos', como sucede más comúnmente en los EC, el diseño será llamado factorial 2x2. La figura 1 muestra un esquema de estos diseños.

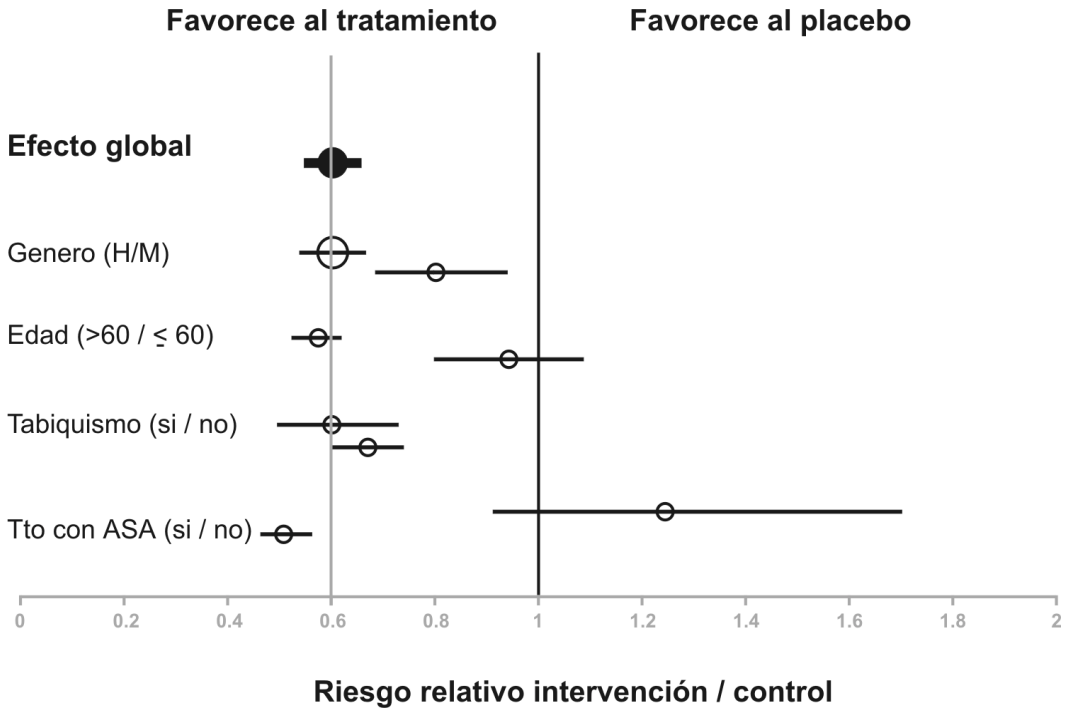

Figura 1. Ejemplo de un análisis de subgrupos en un ensayo clínico hipotético en el cual se prueba una intervención que resulta en un riesgo relativo (RR) de 0.60 respecto al tratamiento control en toda la población del estudio. Las líneas horizontales y los círculos representan el tamaño del efecto y sus intervalos de confianza del 95%. Ver más explicaciones en el texto.

Los EC con diseños no paralelos tienen como ventaja una mayor eficiencia al requerir la mitad de la muestra (diseño cruzado) o probar dos (o más) intervenciones en el mismo EC (factorial). Sin embargo, debe el lector tener en cuenta que en el diseño cruzado se requiere que el curso natural de la enfermedad se mantenga más o menos estable (si no se hubiese realizado ninguna intervención), que la enfermedad no tenga un tratamiento conocido y que el efecto que se mide de las intervenciones en prueba desaparezca (o pueda ser 'lavado') completamente antes de administrar la siguiente. En el diseño factorial se requiere que el efecto (o el mecanismo al que este se atribuye) sea independiente en cada uno de los factores experimentales. Se invita al lector interesado a revisar ejemplos de EC publicados que han utilizado estos diseños.[17, 18]

Los desenlaces a ser comparados en cada uno de los grupos del EC se registran tiempo después de administrar los tratamientos. Aunque en los EC se prevé un tiempo en el que deberían hacerse patentes las diferencias que se consideren clínicamente importantes, un efecto (deseado o no) podría verificarse antes. También puede interesar a los investigadores la determinación del momento en el cual un efecto aparece, alcance su punto máximo o desaparece luego de haber sido evidente. La necesidad de la evaluación de los efectos de las intervenciones a través del período de seguimiento implica establecer en el EC uno o más tiempos de referencia (*time points*). En nuestro ejemplo, el tiempo de analgesia se categorizó en períodos de seis horas durante las 24 horas de seguimiento, lo que permitió a los autores evaluar diferencias en la duración de la analgesia proporcionada por las mezclas a través del tiempo (Tabla 5 del informe). Una razón adicional para el establecimiento de tiempos de referencia en los EC es la de poder advertir diferencias clínica o éticamente importantes antes del final del período de seguimiento del estudio y que eventualmente podrían ameritar modificaciones, incluso la suspensión del EC. Esto es particularmente importante si el estudio ha propuesto la evaluación de desenlaces "duros" ("*hard" outcomes*), como se denomina a aquellos que generan altos costos sociales o económicos (muertes, hemorragias cerebrales, cirugías u hospitalizaciones, etc).

Aquí resulta crítico asegurar que si se hacen evidentes diferencias durante el período de seguimiento pueda tomarse acción inmediata. Por esta razón, en la organización de los EC se designa a un grupo de personas ajenas a los investigadores al cual le son referidos los datos parciales del estudio periódicamente para su monitoreo. El grupo evalúa si las diferencias entre las intervenciones traspasan una barreras de seguridad (*safety boundaries*) preestablecidas, de modo que el estudio pueda ser interrumpido si los beneficios del tratamiento son evidentemente benéficos antes de terminar el estudio (lo que haría antiético no suministrar el tratamiento experimental al grupo control) o, por el contrario, si los efectos adversos del tratamiento experimental hacen antiético continuar su administración en el estudio. Este grupo es usualmente denominado comité de vigilancia de los datos y monitoreo del estudio (*Data Safety and Monitoring Board*, or DSMB). Estas barreras de seguridad buscan beneficiar o proteger a los participantes de efectos tempranos o imprevistos, cuya notificación no debería esperar hasta su observación por los investigadores al final del estudio. Si así lo desea, el lector interesado es dirigido a ejemplos de EC terminados prematuramente por petición de estos comités de vigilancia con base en diferentes consideraciones.[19-21]

Es obvio que a través del desarrollo del estudio, especialmente en los EC con un período de seguimiento grande, se pierda el registro de una porción de los participantes. Estas pérdidas del seguimiento (*drop outs*) pueden ser debidas a muchos factores, entre otros el cambio de residencia, la renuencia activa a continuar en el estudio por efectos secundarios u otras diversas razones, incluso la muerte no registrada de participantes. La razón más frecuente de pérdidas en los EC la constituye el no cumplir el régimen de tratamiento que se está probando, por alguna razón. Al cumplimiento de los tratamientos u otras normas del protocolo durante el transcurso del EC se le denomina adherencia (*compliance*). Rangel y cols[1] informaron seis pérdidas en 104 participantes aleatorizados durante un período de seguimiento de 24 horas; cuatro de ellas se debieron a pérdida de la información y dos a falta de adherencia al protocolo. Una práctica corriente en los EC para disminuir las pérdidas futuras por falta de adherencia es la de administrar el tratamiento experimental por un período de prueba (*run-in period*) dentro del estudio, con el fin de evaluar preliminarmente la tolerancia y el cumplimento de la terapia. Al terminar este período, se procede a hacer la asignación aleatoria de tratamientos en quienes toleren la medicación y han demostrado una buena adherencia al protocolo del estudio.

Al evaluar un EC, el lector debe estar atento al registro de la cantidad y calidad de esas pérdidas, puesto que ellas afectan la validez de las inferencias en el estudio.

Si las pérdidas del seguimiento en un estudio son altas, esto aumentará por si mismo la probabilidad de errores alfa o beta. Si las pérdidas son claramente mayores en uno de los grupos, estas contribuirán a los sesgos de selección al producir desbalance entre los grupos a ser comparados. Cualitativamente, debe observarse en el informe el detalle de las razones por las cuales los participantes se han perdido del seguimiento. Allí puede omitirse alguna razón que afecte en forma sistemática a un grupo de tratamiento. En general, se considera poco aceptable un EC con una pérdida de participantes mayor al 20%, o con pérdidas que provoquen desbalance entre los grupos. Debe aclararse, sin embargo, que pérdidas del seguimiento no es sinónimo de pérdidas de información. Aun cuando los participantes no continúen tomando su tratamiento experimental (o placebo) o no hayan asistido a un control para la evaluación de sus desenlaces, el estudio debe hacer todos los esfuerzos por recoger la mayor información posible de los participantes durante el seguimiento, en el peor de los casos su estado de supervivencia. Se considera inaceptable que no se tenga ninguna información en más del 5% de la población aleatorizada. Frecuentemente los informes de los EC presentan esta información en una figura llamada flujo del estudio (*study flow*). La figura 2 muestra un diagrama de flujo que resume el proceso de conducción de un EC.

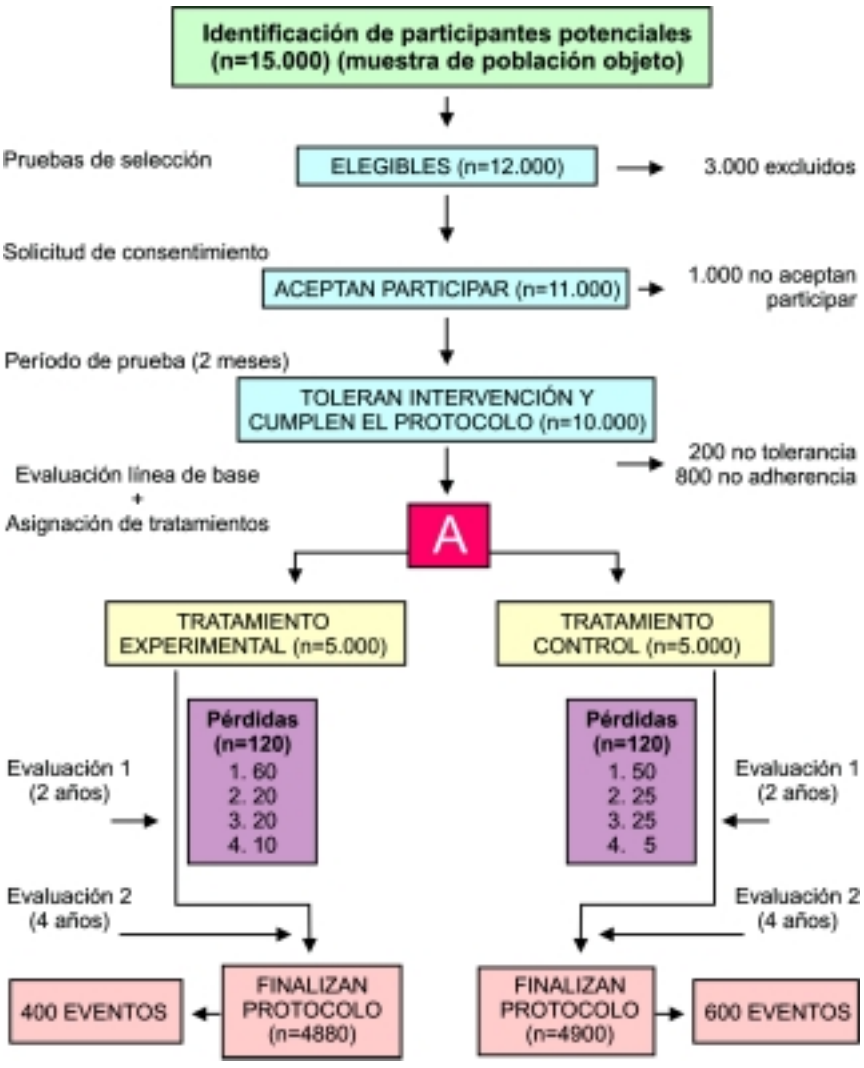

Figura 2. Diagrama que representa el flujo del estudio al que se refiere la figura 1. Se muestran las diferentes etapas del proceso de conducción de un EC y el número de participantes en cada una de estas etapas.A: Aleatorización. En el cuadro que explica las pérdidas del seguimiento se dan cinco razones: 1. Abandono del tratamiento; 2. No asistencia a controles; 3. Abandono por razones médicas; 4. Cambio de residencia; 5. Sin información.

Las pérdidas de un EC estudio debilitan las inferencias que puedan ser tomadas de este. Una estrategia que se ha utilizado para disminuir los sesgos derivados de las pérdidas del seguimiento es la de analizar los desenlaces en toda la población aleatorizada de la cual se tenga información de los desenlaces, no solamente en la que terminó el estudio. Este será un análisis por intención de tratamiento (*intention-to-treat analysis*) basado en la conformación de los grupos señalada por la aleatorización, más que en la situación de haber o no permanecido en el estudio durante todo el período de seguimiento, en el entendido de que esas situaciones (como ocurre también fuera de los estudios) se deben a factores comunes a los dos grupos, excepto por el efecto del tratamiento que está siendo evaluado.

Una manera de manejar los datos perdidos en los EC es la de atribuir un desenlace (usualmente el peor posible) a todos los casos en los que se carece de información sobre los desenlaces. En este tipo de análisis los investigadores buscan la evaluación de la fortaleza (*robustness*) de sus resultados, en el caso que esos participantes perdidos enfrentaran el peor escenario posible. Usando este abordaje, en ocasiones se reporta un análisis del mejor y peor escenario (*best-case, worst-case scenario analysis*).

Otro procedimiento que evalúa la fortaleza de la comparación primaria se da en los EC de gran tamaño, en donde dada la inclusión de un gran espectro de participantes, los investigadores confrontan el resultado de la comparación global en el grueso de los participantes frente a los resultados en diferentes subgrupos de participantes.[17, 22] Esto se conoce como análisis de subgrupos. Aquí se pueden responder preguntas secundarias al establecer si el estimado del tamaño del efecto obtenido incluyendo la totalidad de participantes es similar a través de estos subgrupos. La figura 3 muestra un ejemplo de un EC hipotético en donde se encuentra un riesgo relativo para eventos entre el grupo tratado y el grupo control de 0.60 (IC 95% 0.55, 0.66). Se analiza el tamaño del efecto de acuerdo con la presencia o ausencia de categorías en cuatro factores: género (masculino o femenino); edad (mayor de 60 años, 60 años o menor); tabaquismo (fumador, no fumador/exfumador) o tratamiento concomitante con ácido acetilsalicílico (ASA). El efecto protector del tratamiento experimental en general parece no verificarse en personas menores de 60 años o tratadas con ASA. Este efecto parece ser menor, aunque aun benéfico, en las mujeres que reciben tratamiento.

De este ejemplo resulta obvio que el efecto del tratamiento experimental no es homogéneo en todos estos subgrupos de pacientes. Puesto de otra forma, la interacción entre algunos de estos factores y el efecto del tratamiento es significativa (lo que puede deducirse de la ausencia de superposición entre las líneas del efecto y sus 95% CI de aquellos factores y las categorías de esos factores). A la inversa, no hay evidencia de interacción significativa entre el tabaquismo y el efecto de la intervención. Nótese la relación de este proceso con el de un meta-análisis de ensayos clínicos, en donde el proceso se realiza a la inversa, partiendo de resultados parciales (los EC individuales) para calcular un estimado global del tamaño del efecto.[23] El análisis de subgrupos va del todo a las partes. Al igual que en los meta-análisis, los autores a menudo informan el resultado de la prueba estadística para descartar formalmente si hay heterogeneidad del efecto entre estos subgrupos. En general, todos aquellos análisis en los EC en donde se compara la fortaleza de la comparación primaria del estudio frente a otras comparaciones secundarias (subgrupos particulares o análisis de desenlaces simulando diversos escenarios) se llaman análisis de sensibilidad (*sensitivity analysis*) del resultado.

El establecimiento de diferencias entre los grupos de tratamiento es particular a cada EC, pues estas se dan dentro de su contexto clínico, el marco de su diseño, la precisión estadística dada por el tamaño de su muestra y el error en la medición de los desenlaces.[24] En el diseño de un EC se parte de una pregunta que los autores consideran relevante, a partir de la cual estos preestablecen una diferencia que ellos consideran clínicamente importante para cierto(s) desenlace(s). El establecimiento de esta diferencia es generalmente el resultado de la observación de los autores en su práctica asociado o no al soporte dado por publicaciones previas en el tema (no siempre existentes, especialmente si el ensayo es en realidad una 'novedad'). Rangel y cols[1] basaron el estimado de la muestra en una diferencia esperada para el tiempo de latencia de (al menos) un minuto. Este punto de partida para los autores es su diferencia clínicamente significativa, el umbral por encima del cual los investigadores considerarían se debería sugerir a los usuarios la modificación de su práctica.

En cambio, una diferencia estadísticamente significativa ($p<0.05$ usualmente, pero no necesariamente, de acuerdo con el contexto del EC) puede alcanzarse con diferencias que puedan ser clínicamente irrelevantes. Supongamos que en la evaluación del efecto de dos dosis diferentes de un antihipertensivo, asignamos aleatoriamente tratamiento con 10 mg (grupo "A") y con 50 mg (grupo

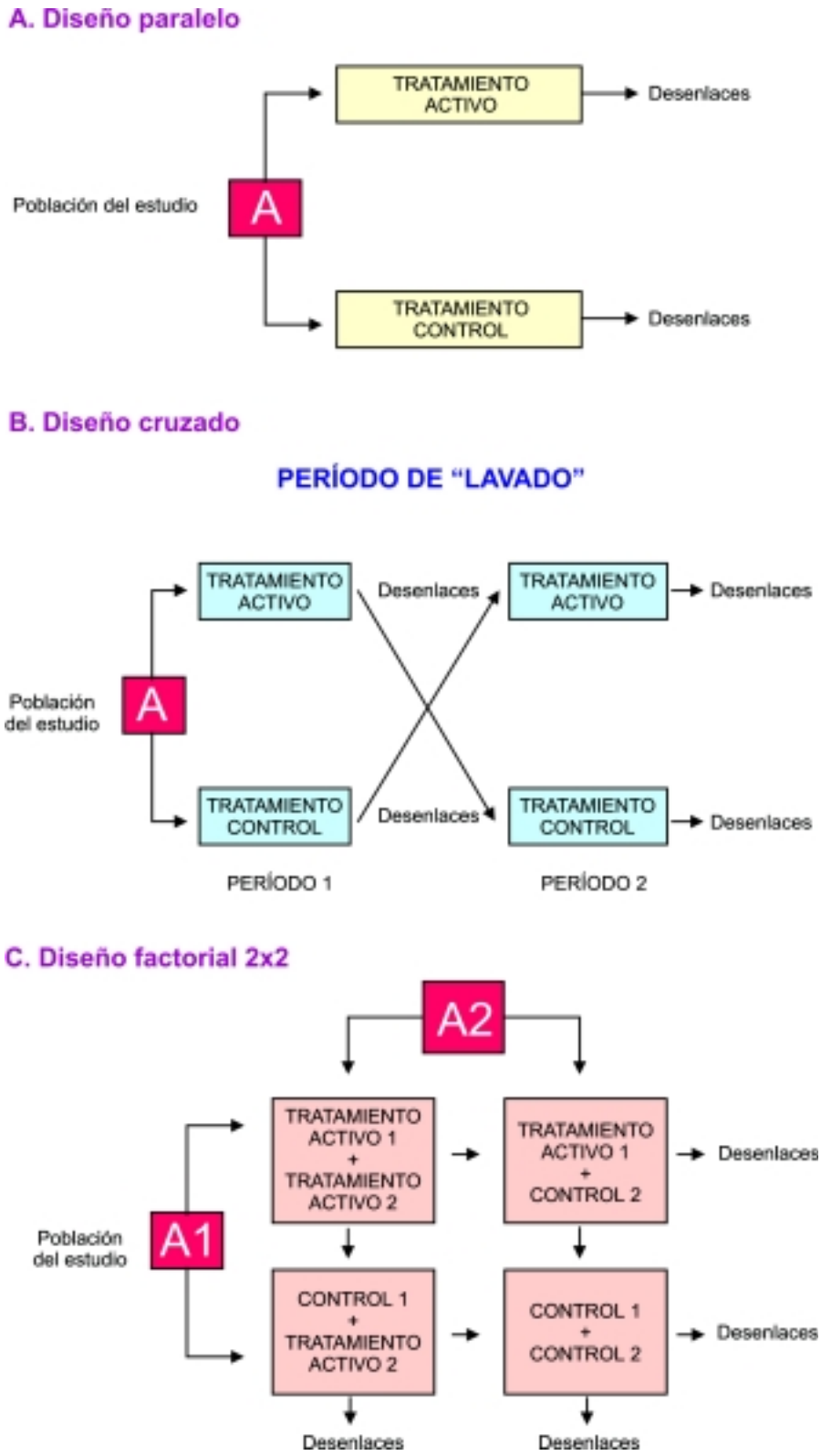

Figura 3. Esquema que muestra tres diseños en los ensayos clínicos. Cada cajón representa un grupo de participantes asignado al mismo tratamiento. A: Aleatorización de la asignación de las interveniones. Ver comentarios en el texto.

"B") a dos grupos de pacientes. Después de un mes de tratamiento, la media de la presión arterial sistólica del grupo A fue 122 mm Hg con una desviación estándar (ds) de 15 mm Hg, mientras que la del grupo B fue 120 (ds 15). Si en el contexto clínico de los autores se considera que la diferencia obtenida de 2 mm de Hg no corresponde a la respuesta terapéutica esperada luego de aumentar cinco veces la dosis del medicamento, entonces esta diferencia no sería clínicamente importante. Sin embargo, si el estudio está hecho en un número suficientemente grande de participantes, esta diferencia puede alcanzar el umbral de 'significancia estadística' ($p<0.05$). La parte izquierda de la tabla 3 muestra el valor de p para esa diferencia mientras el tamaño de los grupos crece. Puede notarse allí que para la misma diferencia, cuando el estudio tiene una muestra de 1.280 o más participantes el valor de p se torna 'significativo'. El valor de p por lo tanto carece de valor por fuera del contexto clínico. En ocasiones puede ser más importante la demostración de que cierta intervención es tan buena como otra en términos de eficacia (pero quizá diferente en términos de costos, facilidad de aplicación, incidencia de efectos secundarios, etc). No se quiere demostrar aquí que A es mejor que B, sino que A y B son equivalentes. Nuestro estudio hipotético pudo haber sido diseñado para probar que el aumento de 10 a 50 mg de ese medicamento no genera diferencias (clínicamente) significativas, luego estas dos dosis serían equivalentes. En estos EC, llamados EC de equivalencia, el análisis estadístico que interesa es el de demostrar que no existen diferencias entre los desenlaces comparados. Se desea por tanto minimizar la probabilidad de error beta. La probabilidad de 'éxito' en evitar el error beta se le denomina poder (1 - la probabilidad de cometer error beta). Usualmente es aceptable que se calcule una muestra para tener un poder de al menos 80% (una probabilidad de error beta del 20%), aunque esto depende del contexto clínico. Los EC de equivalencia suelen diseñarse con un poder de al menos 95%. La parte derecha de la tabla 3 muestra el poder alcanzado cuando se realiza la comparación de la presión arterial alcanzada con las dosis de 10 y 50 mg del antihipertensivo, en caso de que se concluyera una equivalencia clínica en el efecto de estas intervenciones.

Nótese que si se declarase que esas intervenciones fuesen equivalentes, solamente a partir de un número de 1.280 participantes en el estudio podríamos reducir suficientemente la probabilidad de error beta. Rangel y cols[1] señalan la necesidad de al menos 18 pacientes para demostrar diferencias de al menos 1 minuto entre los grupos. Este es el número que obtuvimos (Minitab, professional version 12.22) cuando estimamos el tamaño de la muestra en dos grupos basados en una diferencia de 1 (minuto), teniendo una desviación estándar de 1 y un poder del 80% (el lector puede comparar con las medias y la desviación estándar que fueron obtenidas por el estudio). Estos son los conceptos que el usuario de un EC debería tener a la mano para evaluar la presencia o ausencia de las diferencias que los autores presentan en el informe de un EC.

Tabla 3. Evolución de los valores de p (izquierda) y el poder alcanzado (derecha) para la prueba de la hipótesis de no diferencia entre dos grupos con las características descritas en el texto a medida que el tamaño de la muestra aumenta (centro).

VALOR DE p	TAMAÑO DE LA MUESTRA	PODER (%)
0.8360	10	5.92
0.7683	20	6.96
0.6754	40	9.06
0.5526	80	13.35
0.4003	160	22.12
0.2339	320	39.14
0.0921	640	66.40
0.0172	1280	91.29
0.0007	2560	99.75
0.0000	5120	100

Agradecimientos

Juan C. Villar recibe actualmente una comisión de estudios de la Universidad Autónoma de Bucaramanga.

Summary

Clinical trials (CT) are the basis for generating evidence-based recommendations on therapy. Therefore, it is desirable that the local academic community increases its awareness and skills on critical appraisal of CTs to judge their validity, weight their effect estimates and decide how applicable to practice a CT may become. This paper (first of two parts) outlines the more important concepts

and characterisitics of CTs, while introducing some terms that are often used in their reports. The following issues are discussed: a) The origins and consequences of alpha and beta error in CTs; b) The principle of randomization and how to assess the balance of study groups at the baseline. c) Different ways by which comparisons between study groups can be set up as basis for making inferences.

Key words: Clinical trials, methodology, randomization

Referencias

1. Rangel GW, Chona JE, García LM. Bloqueo peribulbar adicionando opioides en analgesia postoperatoria para cirugía oftálmica. Un ensayo clínico controlado. MEDUNAB 2001; 4:1-8.
2. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 1993; 270:2093-5.
3. Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt G. Users' guides to the Medical Literature. VIII. How to use clinical practice guidelines. B. what are the recommendations and will they help you in caring for your patients? The Evidence-Based Medicine Working Group. JAMA 1995; 274:1630-2.
4. Diccionario Oxford Español-Inglés Inglés-Español. 1 ed. Oxford: Oxford University Press; 1994.
5. Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR. The relation between systematic reviews and practice guidelines. In: Mulrow C, Cook D, eds. Systemic reviews. Synthesis of best evidence for health care decisions. 1 ed. Philadelphia: American College of Physicians; 1998: 55-65.
6. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. BMJ 1991;303: 1499-503.
7. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. Lancet 1995;345: 669-85.
8. Oto A. Magnesium treatment in acute myocardial infarction: an unresolved consensus. Eur Heart J 1999;20: 86-8.
9. Mehta SR, Eikelboom JW, Yusuf S. Risk of intracranial haemorrhage with bolus versus infusion thrombolytic therapy: a meta-analysis. Lancet. 2000;356: 449-54.
10. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 1. Hypothesis testing. Can Med Assoc J 1995; 152:27-32.
11. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? Stat Med 1984; 3:409-22.
12. Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C et al. A cohort study of summary reports of controlled trials. JAMA 1990; 263:1401-5.
13. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994; 272:125-8.
14. Schulz KF, Chalmers I, Altman DG, Grimes DA, Dore CJ. The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals. Online J Curr Clin Trials 1995;Doc No 197:81.
15. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990; 335:149-53.
16. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. CMAJ 1995; 152:169-73.
17. Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, Dagenais G. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. The Heart Outcomes Prevention Evaluation Study Investigators. N Engl J Med 2000; 342:145-53.
18. Venkatesh G, Fallen EL, Kamath MV, Connolly S, Yusuf S. Double blind placebo controlled trial of short term transdermal scopolamine on heart rate variability in patients with chronic heart failure. Heart 1996; 76:137-43.
19. The Cardiac Arrhythtmia Suppression Trial (CAST) investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 1989; 321:406-12.
20. Englund JA, Baker CJ, Raskino C, McKinney RE, Petrie B, Fowler MG et al. Zidovudine, didanosine, or both as the initial treatment for symptomatic HIV-infected children. AIDS Clinical Trials Group (ACTG) Study 152 Team. N Engl J Med 1997; 336:1704-12.
21. Randomized trial of aspirin, sibrafiban, or both for secondary prevention after acute coronary syndromes. Circulation 2001; 103:1727-33.
22. Connolly SJ, Kerr CR, Gent M, Roberts RS, Yusuf S, Gillis AM et al. Effects of physiologic pacing versus ventricular pacing on the risk of stroke and death due to cardiovascular causes. Canadian Trial of Physiologic Pacing Investigators. N Engl J Med 2000; 342:1385-91.
23. Villar JC. Más y mejores revisiones sistemáticas para un mejor resultado en nuestra práctica. MEDUNAB 2000; 3:47-54.
24. Sackett DL, Cook RJ. Understanding clinical trials. BMJ 1994; 309:755-6.