PUESTA AL DÍA

Métodos de investigación en cardiología clínica (VII)

Estudios experimentales en cardiología

Javier Muñiz[a,b], Darwin R. Labarthe[c], Rafael Juane[b,d] y Alfonso Castro Beiras[b,e]

[a]Hospital Juan Canalejo. Instituto de Ciencias de la Salud. Universidad de La Coruña. La Coruña.
[b]Asociación de Estudios Biomédicos de Galicia (BIOMEGA). La Coruña.
[c]School of Public Health. University of Texas-Health Science Center. Houston. Texas. EE.UU.
[d]Sanatorio Quirúrgico Modelo. La Coruña.
[e]Servicio de Cardiología. Hospital Juan Canalejo. La Coruña.

cardiología/ enfermedades cardiovasculares/ ensayos clínicos/ tamaño muestral

En este artículo se discuten los diseños experimentales en investigación clínica. Se utilizan como hilo conductor ejemplos del área de la cardiología, en la medida de lo posible siempre el mismo. Se ha buscado una óptica diferente de la habitual en la discusión de las características generales de los diseños experimentales y más específicamente de los ensayos clínicos y se han abordado aspectos de los mismos que habitualmente escapan de la consideración de los clínicos por su carácter marginal.

Se inicia discutiendo qué características son diferenciales de los ensayos clínicos respecto a otros diseños y el tipo de pregunta a que responden aquéllos. A continuación se discuten aspectos como aleatorización y tipos (simple, por bloques, estratificada, prealeatorización), tipos de variables de evaluación de la respuesta, enmascaramiento y problemas para mantenerlo con determinados diseños, tamaño muestral, etc.

Se hace una breve mención de dos casos particulares: el diseño factorial y el diseño cruzado y se discuten sus puntos fuertes y débiles. Asimismo, se discute otro diseño experimental como es el ensayo comunitario y se aportan ejemplos. El artículo finaliza con la discusión de aspectos de criterio como cuándo parar un ensayo o a quién son aplicables sus resultados y se sugieren puntos a tener en cuenta en esas tomas de decisión.

EXPERIMENTAL STUDIES IN CARDIOLOGY

Experimental designs in clinical investigation are discussed in this article. Guideline examples have been used in the area of Cardiology using always the same one only one whenever possible. We have looked for a different perspective from what is generally used in the discussion of the general characteristics of experimental designs, and more specifically of clinical trials and we deal with the aspects of clinical trials which are usually ignored due to their marginal character.

We also discuss those characteristics which differentiate clinical trials in respect to other designs and types of questions which are answered by clinical trials. And we finally discuss various aspects such as randomization and its various types (simple, block, stratified, pre-randomized) and variable types of evaluating the answers, masking and the problems in its maintenance, with certain kinds of designs, sample size, etc. There is a brief mention of two particular cases: factorial and cross over designs are both discussed, mentioning their strong and weak points. Likewise, we discuss community trials as another experimental design and examples are provided.

Finally, we discuss aspects of criteria: such as, When to stop the trials? or Who are the results applicable to?, and we suggest points to take into consideration when these decisions are made.

(*Rev Esp Cardiol* 1997; 50: 268-277)

INTRODUCCIÓN

Experimentar es, en las ciencias fisicoquímicas y naturales, hacer operaciones destinadas a descubrir, comprobar o demostrar determinados fenómenos o principios científicos[1].

Correspondencia: Dr. J. Muñiz.
Instituto de Ciencias de la Salud. Hospital Marítimo de Oza.
Pabellón n.º 6. 15006 La Coruña.

Dentro de esta definición tan genérica, en lo que se refiere a ciencias de la salud se incluyen una amplia variedad de estudios, en los que el sujeto de estudio puede ser un aparato, unas células, un animal o una persona. Incluso, conceptualmente, como veremos más adelante, una población.

Este trabajo se refiere a los experimentos que incluyen personas. Más concretamente a aquellos en los que la unidad de intervención (y de evaluación de la respuesta) es el individuo. Se hará una breve mención a los ensayos comunitarios sin entrar en detalles, pero el núcleo del trabajo son los ensayos clínicos, preferentemente los de prevención secundaria. Se asume que el lector medio de la revista está acostumbrado a muchos de los conceptos inherentes a los ensayos clínicos, por lo que este artículo, si bien necesariamente incluirá la exposición de aspectos generales de este tipo de diseño, intentará buscar una óptica o punto de vista alejado del habitual y abordar aspectos marginales de los ensayos clínicos, intentando incorporar elementos de reflexión y discusión relativos a estos aspectos. Adaptándose a las recomendaciones de los editores invitados[2], se ha pretendido conjugar lo anterior con la realización de un artículo de divulgación metodológica, centrado en aquellos aspectos o áreas de interés del cardiólogo clínico e intentando abordar dudas o problemas frecuentes con los que el investigador (o el lector) se puede encontrar al iniciar (o leer) un determinado trabajo experimental, con el objetivo último de que estas reflexiones resulten útiles y prácticas para el lector medio de la revista[3]. Se renuncia explícitamente a aspectos de cálculos estadísticos formales para centrarse, cuando se consideró necesario, en aspectos de lógica científica[3]. Siguiendo la estructura común a la serie, siempre que se consideró necesario un ejemplo, se utilizó, en la medida de lo posible, el mismo ejemplo, en este caso el β-HAT (Betablocker Heart Attack Trial)[4,5], un ensayo de prevención secundaria bien conocido por los cardiólogos, que ha cambiado la práctica clínica y con los suficientes años de antigüedad como para que los conocimientos que ha aportado estén integrados en la práctica cardiológica. De hecho, los resultados principales de este ensayo se publicaron en 1982, fecha en la que una gran proporción de los cardiólogos en activo en la actualidad no habían terminado su período de entrenamiento. Por tanto, para una parte importante de los cardiólogos, ese saber siempre estuvo ahí. Otra parte importante habrá vivido todas las dudas previas a la aparición de este ensayo y otros similares.

¿QUÉ CARACTERÍSTICAS DE UN ENSAYO CLÍNICO HACEN QUE SEA UN ENSAYO CLÍNICO?

Bradford Hill definió el ensayo clínico como «un experimento diseñado cuidadosa y éticamente con el objetivo de responder a alguna pregunta formulada con precisión»[6]. Años después, Bulpitt, en su libro *Randomized Controlled Clinical Trials*[7], parafraseando esta definición, define un ensayo clínico aleatorizado y controlado como «un experimento diseñado cuidadosa y éticamente que incluye la provisión de controles adecuados y apropiados mediante un proceso de aleatorización con el objetivo de poder responder a preguntas formuladas con precisión».

El β-HAT, como se indica en el resumen y en el apartado de material y métodos[4,5], es un ensayo clínico *multicéntrico, aleatorizado, doble ciego y controlado con placebo.* El lenguaje utilizado en las revistas científicas, y en menor medida en los libros, pretende, y suele conseguirlo, ser lo más sintético posible, por lo que la mera adición de todos los adjetivos que siguen al sustantivo «ensayo clínico» sugeriría que puede haber ensayos clínicos a los que les falte alguna o todas las características que califican a este ensayo en particular: multicéntrico, aleatorizado, doble ciego y controlado con placebo. Pero, ¿es esto cierto?

Dejando de lado por el momento el que sea multicéntrico, pues sólo hace referencia, aunque no es poco, al hecho de que se realiza en varios centros, las otras características califican al ensayo clínico idóneo, pero quizá no sean necesarias para que un determinado estudio entre en la categoría de ensayo clínico. Atendiendo a la pregunta del encabezado: ¿qué características convierten a un determinado estudio en ensayo clínico?, la tabla 1 resume una propuesta de las características necesarias y/o singulares de este tipo de diseño. Discutiremos una por una estas características en el convencimiento de que, al recoger esta tabla opiniones de los autores y no dogmas, para muchas de ellas no habrá unanimidad. El signo (+) indica que esa característica es necesaria o única, y el signo (–) que no lo es. Se indican con (+/–) aquellas características para las que no hay uniformidad de criterio entre diferentes autores. La tabla, en

TABLA 1
Características necesarias y únicas de un ensayo clínico

	Necesario	Único
Una intervención *elegida por el investigador*	+	+
Observación prospectiva, seguimiento	+	–
Evaluación de una respuesta o evento	+	–
Grupo de control o de comparación concurrente	+/–	–
Colocación aleatoria, o aleatorización, tratamiento y control	+/–	+
Controles con placebo (no «no tratados»)	+/–	+
Ciego, o no revelar la asignación a tratamiento a uno o más niveles, por ejemplo, «doble ciego»	+/–	–
Consentimiento informado	+	–

cualquier caso, es meramente orientativa y, en particular en lo relativo a características necesarias, es útil reflexionar en cada caso más allá de la mera terminología de si lo que estamos leyendo (y queriendo aplicar) es un ensayo clínico o no, para considerar en qué medida afecta la presencia o no de una determinada característica a la credibilidad del trabajo (entendida ésta de manera genérica) y, en último extremo a su validez y utilidad. No obstante, hay varias características sobre las que existe un consenso generalizado: debe haber una intervención elegida por el investigador y debe medirse el efecto que esta intervención produce en los intervenidos cierto tiempo después (este tiempo puede oscilar entre unos segundos y muchos años, pero la intervención siempre precede a la evaluación de la respuesta) lo que lo convierte, por definición, en un estudio prospectivo. Esto, conceptualmente, no precluye el hecho de que uno o más brazos pueda ser construido mediante una cohorte histórica: se utilizarían en este caso datos ya recogidos pero la inclusión de los pacientes sería en función de la intervención que se les administra y no del desenlace que tienen, por lo que no se pierde el carácter de prospectivo.

TIPO DE PREGUNTA A LA QUE RESPONDEN LOS ENSAYOS CLÍNICOS

El tipo de pregunta genérica a que responden los ensayos clínicos es el de ¿la modificación de un factor altera el curso de la enfemedad? Esta formulación genérica se puede concretar en dos grandes grupos de preguntas. Por una parte, en la pregunta del tipo de ¿desaparecerá la complicación Y o disminuirá su severidad al tratar a aquellos pacientes que presentan Y con A en comparación con B? U otro tipo de preguntas como es ¿el tratamiento con A producirá una reducción en la incidencia de la complicación Y en pacientes con la condición X en comparación con pacientes similares tratados con B? En el caso concreto del β-HAT el objetivo principal o la pregunta a la que pretende responder es de este último tipo y se formula a los autores de la siguiente manera[4]: ¿puede reducirse la mortalidad (Y) después de un infarto agudo de miocardio (X) mediante tratamiento a largo plazo con propranolol (A) en comparación con placebo (B)?

La primera formulación puede aplicarse, por ejemplo, a cualquier ensayo clínico con un fármaco antihipertensivo en el que la evaluación de la respuesta de interés sea precisamente la modificación de la presión arterial. Siguiendo la tabla 1, la intervención elegida por el investigador es en este caso la administración de propranolol después del infarto agudo de miocardio. Se sigue prospectivamente a estos sujetos y se evalúa una respuesta que en este caso la principal evaluación es si el sujeto o sujetos se mueren o no.

ALEATORIZACIÓN

El objetivo de realizar una aleatorización es conseguir que los diferentes grupos sean comparables u homogéneos, evitar el sesgo del investigador en la asignación de enfermos a tratamientos y garantizar que los tests estadísticos tendrán valores de significación estadística válidos[8]. De hecho, la aleatorización pretende no sólo que sean comparables u homogéneos para aquellas cosas o aquellas variables que medimos sino que lo sean para una infinidad de otras condiciones o variables que no sólo no hemos medido sino que cabe la posibilidad de que ni siquiera sepamos de su existencia pero que potencialmente puedan tener una importancia capital en el desenlace que nosotros evaluamos.

La aleatorización consiste en asignar de manera aleatoria a los pacientes a cada grupo de tratamiento. Vemos, por tanto, que el investigador decide menos de lo que parece y deja en manos del azar el colocar a los pacientes en un grupo o en otro. La aleatorización intenta evitar también que el investigador desempeñe un papel en asignar determinado tratamiento a determinados grupos de pacientes e incluso de manera bien intencionada (por ejemplo, aquellos en los que él cree que se van a beneficiar más de un determinado tratamiento asignarlos a ese tratamiento y no al otro).

El objetivo de que ambos grupos sean «comparables» no siempre está garantizado. De hecho, cuanto menor sea el número de pacientes incluido en el ensayo, más probable es que haya diferencias entre el grupo placebo y el grupo control. Estas diferencias, en ensayos particularmente pequeños, pueden incluso no ser estadísticamente significativas pero sí tener relevancia clínica. Un ejemplo de esta afirmación puede ser el de un ensayo clínico con 12 pacientes en el grupo de tratamiento A y 12 en el grupo de tratamiento B, la aleatorización puede haber colocado a 4 varones en un grupo y 8 en el otro. Esta diferencia no alcanza el nivel de significación estadística de $p < 0,05$ pero es difícil sostener que ambos grupos son similares en lo que respecta a la proporción de varones en cada uno[7]. Como decíamos antes, para aquellas variables de las que conocemos su efecto sobre el desenlace que nosotros medimos o sobre efectos de los tratamientos que nosotros aplicamos, y que además las consideramos y medimos, el problema no es excesivamente grave. Esto es así porque, como veremos más adelante, en el apartado de diseño se puede estratificar por posibles variables de confusión (por un número limitado de ellas) para garantizar que ambos grupos están equitativamente distribuidos a ese respecto. Si no se ha hecho esto, y nos encontramos con que hay diferencias entre ambos grupos, incluso a la hora del análisis, al tener recogida esa variable, pueden ajustarse de alguna manera estos desequilibrios. El problema surge con toda la infinidad de potencia-

les variables de confusión, muchas de las cuales ni siquiera sabemos que existen ni el papel que desempeñan y para las que es imposible tanto estratificar en la fase de diseño como ajustar en la fase de análisis. Cuando en la aleatorización, por muy bien que se haya hecho, no existen suficientes indicios como para sospechar que ha tenido éxito, los resultados del ensayo o más bien su aplicabilidad o la credibilidad que les atribuimos está comprometida. Los indicios para valorar si la aleatorización ha tenido éxito deben buscarse por una parte en su descripción en el apartado de material y métodos y, por otra, en lo que debería ser la primera tabla de un ensayo aleatorizado: la tabla en la que se comparan las características basales de los dos grupos. En nuestro ejemplo, en la primera tabla del informe principal[5] se comparan más de cuarenta variables importantes, que incluían, entre otras, características demográficas, peso, presión arterial, colesterol sérico, consumo de cigarrillos, antecedentes personales, características del infarto, medicación, etc. Sólo se observaron diferencias, no de gran magnitud, en dos variables. Esto fue así en un ensayo clínico de gran tamaño, con casi dos mil pacientes en cada grupo, lo que se interpretó por los propios autores como una «comparabilidad entre ambos grupos en conjunto excelente».

Aleatorización simple

En la aleatorización simple lo que se realiza en esencia es que cada vez que un paciente entra en el ensayo se tira una moneda y se coloca en uno o en otro tratamiento (en el ensayo más sencillo con sólo dos grupos de tratamiento) dependiendo de que salga cara o cruz. Alrededor de la mitad de pacientes serán asignados al tratamiento A y la otra mitad al tratamiento B. De manera práctica, para solucionar problemas logísticos, toda la secuencia de asignación se realiza antes del inicio del estudio bien con un generador de números aleatorios o utilizando una tabla de números aleatorios y se mantiene fuera del alcance del investigador (en el centro coordinador o mediante otro sistema de ocultación) para evitar que éste conozca a qué tratamiento se asignará el próximo paciente que se incluya en el estudio.

La tabla de números aleatorios citada aparece como apéndice en cualquier libro básico de bioestadística y consiste en una página (o más) cubierta con los dígitos del 0 al 9 dispuestos en filas y columnas que se han creado mediante un proceso que garantiza que todos los dígitos (del 0 al 9) tienen igual probabilidad de ser elegidos en cualquier selección[9]. Los números no tienen ningún orden identificable y, de hecho, el aspecto de la secuencia de dígitos partiendo desde cualquier punto y yendo en cualquier dirección es caótico. En ensayos pequeños, este sistema de aleatorización simple puede ocasionar que haya un desequilibrio de cierta importancia en el tamaño de los dos grupos, lo que no sólo suscita dudas de la comparabilidad de los dos grupos sino que disminuye la eficiencia del estudio. Debido a esto, con frecuencia se realiza la aleatorización por bloques.

Aleatorización por bloques

Se describió hace años[6] y garantiza que en ningún momento del proceso de aleatorización el desequilibrio en el número de individuos en cada grupo será grande y que en ciertos puntos de este proceso el número de sujetos en cada grupo será idéntico[8]. De nuevo, con el caso más sencillo de sólo dos tratamientos (A y B), si se desea aleatorizar con igual probabilidad a ambos grupos (éste es el caso casi siempre), dentro de cada bloque del tamaño que se fije (un número par: 4, 6, 8, etc.) se asigna la mitad de los pacientes a A y la mitad a B. Se aleatoriza el orden en que se asignan los tratamientos en cada bloque y se repite el proceso tantas veces como sea necesario. Por ejemplo, si el tamaño del bloque es 4, tenemos la garantía de que cada cuatro sujetos aleatorizados el tamaño de los dos grupos es idéntico. En este caso, el posible orden de asignación a A o B es uno de los seis siguientes: AABB, ABAB, ABBA, BBAA, BABA y BAAB. Se elige al azar una de las seis secuencias y cuando ya se han incluido cuatro participantes se repite el proceso. Obviamente, si todos los bloques son de longitud = 4, cabe la posibilidad de que, si se desvela el grupo a que pertenecen los dos primeros sujetos de un grupo, se descubra automáticamente el grupo a que se asignarán los dos siguientes (si AA, debe seguir BB, si BAB, debe seguir A, etc.). Para evitar esto se requiere que el tamaño de los bloques varíe, a su vez, en una secuencia aleatoria.

Aleatorización estratificada

Se realiza para garantizar que los grupos de tratamiento no difieren con respecto a una o varias variables pronósticas de interés. Se forman estratos a partir de las características iniciales, por ejemplo, sexo, edad, nivel de colesterol, etc. y después se realiza la aleatorización dentro de cada estrato. El número de estratos que se forma es el resultante de multiplicar el número de grupos en cada variable por la que queremos estratificar. Por ejemplo, dos sexos, cuatro grupos de edad y tres grupos según nivel de colesterol basal producirían 24 ($2 \times 4 \times 3$) grupos de sexo-edad-colesterol. Como es obvio, para poder aleatorizar de manera estratificada es necesario disponer de información respecto a la variable por la que queremos estratificar con anterioridad a la aleatorización.

Dentro de cada uno de los grupos o estratos se puede realizar la aleatorización simple, pero, generalmente, a fin de evitar desequilibrios en el número de indi-

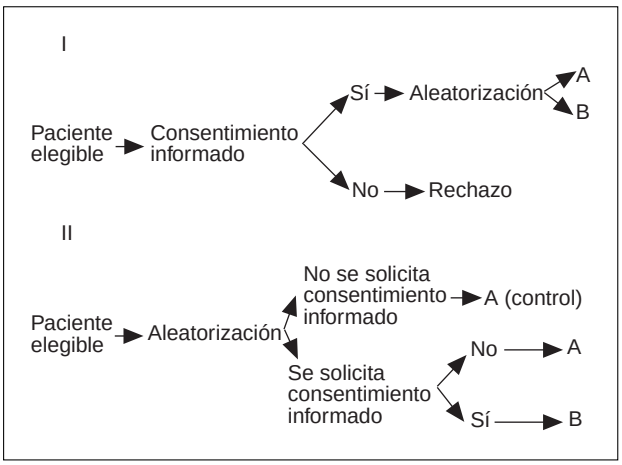

Fig. 1. Secuencia tradicional (I) y alternativa (II) para inclusión y asignación de tratamientos en un ensayo clínico. El tratamiento A se refiere a «la mejor» terapia disponible.

Tratamiento B	Tratamiento A: Sí	Tratamiento A: No*	
Sí	A + B	B	B
No*	A	—**	No B
	A	No A	

*Fig. 2. Diseño básico de un ensayo clínico factorial. *Si es factible, los pacientes reciben placebo (no A, placebo de A; no B, placebo de B y no A, no B placebo de ambos); **no A, no B.*

viduos asignados a cada tratamiento en los grupos, se realiza la aleatorización por bloques.

En nuestro ejemplo, el β-HAT se consideró que, siendo un ensayo multicéntrico y a pesar de todos los procedimientos de normalización que se siguieron, el centro de procedencia del paciente (hospital) podía ser un factor pronóstico de importancia, por lo que se estratificó por centro clínico. En este ensayo participaron treinta y un centros y se consideró importante no sólo esa estratificación sino que la aleatorización garantizara que aproximadamente dentro de cada hospital o centro participante la proporción de enfermos con propranolol y con placebo fuera aproximadamente la misma. Para esto se hizo una aleatorización por bloques en la que la longitud de los mismos variaba aleatoriamente entre 4, 6, y 8[5]. Esto, como se describió con anterioridad, se utilizó como mecanismo de seguridad adicional para evitar que el investigador conociera a qué grupo sería asignado el próximo paciente incluido en el ensayo. Otra situación en que la estratificación puede ser deseable es en el caso de que pueda considerarse una terminación temprana en un subgrupo y continuación de la intervención en otro. Por ejemplo, la estratificación según nivel de presión arterial en el momento de entrar en el ensayo ofrecería la posibilidad de aleatorización independiente para cada estrato y permitiría diferentes tiempos de seguimiento en cada uno sin comprometer el diseño. Este fue el caso de los ensayos de tratamiento de la presión arterial de la Administración de Veteranos[10,11].

Aleatorización tradicional frente a prealeatorización

En lo relativo a aleatorización, el diseño típico de un ensayo clínico sigue la secuencia recogida en la parte superior de la figura 1. Este esquema es el tradicional y el que con mayor frecuencia nos encontraremos en los ensayos clínicos de cardiología. Sin embargo, hay varios aspectos que comprometen la participación en ensayos clínicos tanto de pacientes como de investigadores. Una razón aducida es que a pesar de que en realidad no se disponga de evidencia científica para apoyar un tratamiento u otro (¡imprescindible para que el ensayo esté justificado y sea ético!), los médicos pueden sentirse incómodos en reconocer al paciente que en realidad no saben qué tratamiento es mejor. Por otra parte, particularmente en enfermedades con mal pronóstico y pocas posibilidades terapéuticas, la aparición de un tratamiento prometedor levanta unas expectativas en el paciente, dispuesto a probar «lo que sea». Dejar la decisión de si recibir o no el nuevo y prometedor tratamiento a si sale cara o cruz (¡o de manera más esotérica a la tabla de números aleatorios o al ordenador!) es algo que puede no satisfacer ni al paciente ni al investigador. Para algunos, todos estos requisitos necesarios para el desarrollo de un ensayo clínico tradicional comprometen en cierta medida la relación médico-paciente, y hacen este diseño poco deseable[12].

Para obviar estos problemas, se ha propuesto un diseño alternativo, que se recoge en la parte inferior de la figura 1. En resumen, se aleatoriza a los pacientes elegibles a uno de los dos grupos siguientes: mejor tratamiento disponible hasta la fecha (A) o tratamiento experimental (B). A los que se les asigna A, se les trata sin requisito adicional. A los asignados al grupo B se les explica todas las posibilidades terapéuticas en su caso particular y se les solicita consentimiento informado para ser tratados con el tratamiento experimental (al que el azar le asignó). Si aceptan, se les trata con B y si no aceptan con A. En el análisis de los resultados se comparan los grupos según la aleatorización, independientemente del tratamiento recibido. Este último hecho tiene al menos dos efectos: disminuye la eficiencia del estudio y desagrada a los clínicos. Además, el diseño en su conjunto tiene problemas éticos que han cuestionado su utilidad[13,14].

Respecto al primero de los puntos, en este diseño la eficiencia del estudio es función de la proporción de individuos que aceptan el tratamiento experimental (P) de los que se les propone. La eficiencia de este diseño en comparación con uno tradicional (fig. 1-I) es

P^2, asumiendo que en ambos diseños la mitad de los pacientes son aleatorizados a cada tratamiento pues no hay ventaja en asignar un número desproporcionado de sujetos a ninguno de los grupos. La eficiencia de P^2 significa, por ejemplo, que si el 90% de los pacientes aceptan el tratamiento B (P = 0,9), en este diseño se necesitan 100 pacientes en comparación con 81 pacientes (100 P^2) en un diseño tradicional para disponer de la misma capacidad de detectar diferencias entre tratamientos[12]. Esta pérdida de eficiencia, a juicio del autor, puede verse compensada con el mayor número de pacientes incluido en el estudio.

Respecto al segundo de los puntos, no es fácilmente asumible por el clínico el hecho de que estemos considerando como tratados con B sujetos que sabemos positivamente que han sido tratados con A. El tipo de reparo es de un perfil similar al que se encuentra en la discusión entre los análisis «por intención de tratar» y «por adherencia al protocolo»[7], con la que habitualmente está más familiarizado el clínico. No obstante, sin entrar en consideraciones adicionales, hacer un análisis por tratamiento recibido en lugar de tratamiento asignado puede ocasionar un sesgo importante si la disponibilidad a aceptar un determinado tratamiento está relacionada con el pronóstico[13].

En cuanto a los problemas éticos de la formulación original de Zelen[12] se centran, fundamentalmente, en que no se informa a la mitad de los pacientes de que participaron en un ensayo clínico[14], por lo que no ha tenido una aceptación generalizada. Una adaptación de este procedimiento consiste en prealeatorizar y solicitar el consentimiento a los dos grupos, pero ya conociendo el tratamiento que se les ha asignado[15]. El efecto que sobre el poder del estudio tiene el que puedan no aceptar pacientes de ambos grupos y el «factor de inflación» que este rechazo tiene sobre el tamaño muestral para mantener un determinado poder se recogen en las tablas 2 y 3. Se observa que en determinadas circunstancias puede ser de gran magnitud.

Si bien estas alternativas a la aleatorización tradicional son más frecuentes en ensayos de cáncer, también nos encontramos con ejemplos en el área cardiológica. Este diseño es particularmente deseable cuando existen enormes diferencias entre las intervenciones propuestas, no en el desenlace esperado a priori (podría comprometer la realización de un ensayo clínico), sino en la intervención en sí. Este es el caso, por ejemplo, de los ensayos en los que se comparan tratamientos médicos con tratamientos quirúrgicos. Como es obvio, no pueden aplicarse estos sistemas de aleatorización en aquellos ensayos en los que el mantenimiento del doble ciego sea crucial. No obstante, en determinadas preguntas de gran interés en cardiología no es crucial el que sea doble ciego. Este es el caso cuando la evaluación de la respuesta se hace con una variable de las denominadas «duras», en las que la subjetividad desempeña poco o ningún papel, en oposición a las variables «blandas», en las que la subjetividad puede tener mucha importancia. Un ejemplo típico de variable dura es la muerte, mientras que variable blanda puede ser la sensación de bienestar. Entre ambos extremos se sitúan la mayor parte de las variables de interés en cardiología: reinfarto, necesidad de reintervención, reestenosis, etc. No obstante, incluso cuando la pérdida del enmascaramiento no es una preocupación debido a la elección de un evento final «duro», permanece el riesgo de diferencias sistemáticas entre los grupos en otros tratamientos, comportamiento u otros factores. El coste potencial de un sesgo de este tipo u otras razones que reducen la credibilidad de los resultados pesan en contra de la elección de este tipo de diseños, lo que los hace de uso infrecuente.

Independientemente del tipo de aleatorización utilizado, en determinadas situaciones el mantenimiento del enmascaramiento es imposible en gran parte o to-

TABLA 2
Poder de función de la respuesta real al tratamiento y de la tasa global de rechazo*

Tasa verdadera de respuesta**		Análisis basado en el tratamiento recibido	Poder				
			Análisis basado en el tratamiento asignado. Tasa de rechazo				
P_A	P_B		0,05	0,10	0,15	0,20	0,25
0,10	0,30	0,93	0,86	0,76	0,63	0,49	0,35
0,20	0,40	0,84	0,75	0,64	0,52	0,39	0,28
0,30	0,50	0,79	0,69	0,58	0,46	0,35	0,25
0,40	0,60	0,77	0,67	0,56	0,45	0,34	0,24

*Basado en un tamaño muestral de 100 pacientes en cada rama de tratamiento, un test de dos colas y un error $\alpha = 0,05$; **P_A y P_B son las probabilidades de responder a los tratamientos A y B; tomada de referencia 11.

TABLA 3
«Factor de inflación» de tamaño muestral para diferentes tasas globales de rechazo en estudios prealeatorizados

Tasa de respuesta verdadera	Factor de inflación
0,02	1,09
0,05	1,23
0,10	1,56
0,15	2,04
0,20	2,78
0,25	4,00
0,30	6,25
0,35	11,11
0,40	25
0,45	100
0,50	–*

*Si la mitad de los pacientes de cada rama rechazan el tratamiento asignado aleatoriamente y reciben otro, es imposible determinar diferencias en el efecto del tratamiento independientemente del tamaño de la muestra; tomada de referencia 11.

dos los niveles, como es el caso de la comparación entre tratamiento médico y quirúrgico, lo que nos remite de nuevo a la tabla 1 y a la relatividad de la «necesidad» de esta condición. Un ejemplo reciente en que se utilizó un sistema de prealeatorización fue en un ensayo que comparaba la angioplastia con la cirugía de revascularización en pacientes con lesión única de la descendente anterior y fracción de eyección normal[16].

Otros procedimientos propuestos de asignación de pacientes a uno u otro tratamiento en un ensayo clínico no se discuten por ser infrecuentes en el campo de la cardiología, y más propios de la investigación en cáncer[7,8].

OTROS DISEÑOS

Existen otras variantes de ensayos clínicos. Por su importancia en cardiología se hace una breve mención a dos de ellas: el diseño factorial y el diseño cruzado.

Diseño factorial

Los ensayos clínicos son estudios experimentales que habitualmente llevan aparejado un enorme coste tanto desde el punto de vista del esfuerzo de los investigadores y participantes como coste económico. El diseño factorial intenta, en el caso más sencillo, evaluar mediante un solo experimento dos intervenciones comparadas con control. El diseño básico se recoge en la figura 2. En el mismo, para la intervención A se comparan los grupos A frente a no A (algunos tratados con B tanto en un grupo como en otro) y para la intervención B se comparan los grupos B frente a no B (algunos tratados con A tanto en un grupo como en otro). El realizar dos experimentos o responder a dos preguntas probablemente muy lejanas una de la otra mediante un solo experimento es algo que, dado el enorme coste comentado, resulta atractivo. Sin embargo, una de las limitaciones que tiene es la de la posible interacción entre las dos intervenciones que se realizan, por lo que está particularmente indicado cuando podemos asumir con cierta seguridad que no existe interacción entre las dos intervenciones propuestas. Un ensayo reciente de diseño factorial en el área de la cardiología fue el Physicians Health Study. Este ensayo incluyó a más de 22.000 participantes y se diseñó para probar dos hipótesis de prevención primaria: comprobar si dosis bajas de aspirina en días alternos reducirían la mortalidad por enfermedad cardiovascular, y por otro lado, si la ingesta de β-caroteno en días alternos disminuiría la incidencia de cáncer[17]. Es muy recomendable que se pueda asumir ausencia de interacción porque si no es así se debería estudiar si existe o no entre las dos intervenciones y el obligarnos a eso supone un mayor tamaño muestral ya que el poder para hacer pruebas de interacción es menor que el poder para probar el efecto principal de una determinada intervención[8].

Un ejemplo muy actual en este apartado es el del Antihypertensive Lipid Lowering Heart Attack Trial (ALLHAT) en el que se están estudiando múltiples ramas con diferentes agentes antihipertensivos y dos ramas de tratamiento hipocolesterolemiante (no publicado).

A pesar de lo atractivo que resulta este tipo de diseño, antes de tomar la decisión de usarlo deben también tenerse en cuenta la complejidad añadida y el impacto que puede tener en la adherencia al tratamiento, la selección de pacientes, etc.

Diseño cruzado

El diseño cruzado (*cross-over design*) es una variante especial del ensayo clínico en el que cada paciente es su propio control.

En el caso más sencillo con dos tratamientos A y B, a una serie de pacientes (elegidos aleatoriamente) se les daría primero tratamiento A y después tratamiento B, y a otro grupo de pacientes primero el tratamiento B y después el tratamiento A.

Para poder utilizar este tipo de diseño es imprescindible que los efectos de la intervención o, más concretamente, de la primera fase de la intervención, sea cual sea la que le ha tocado a ese paciente en concreto, no perduren en el momento de la segunda intervención. Uno de los atractivos que tiene es que al ser cada paciente su propio control se disminuyen aspectos de variabilidad entre grupo de pacientes y grupo de controles. Como es obvio, es poco aplicable en el caso de que la duración de la intervención o el tiempo hasta la evaluación de la respuesta desde la intervención sea un tiempo prolongado. Está más indicado cuando la intervención es breve y la evaluación de la respuesta es poco tiempo después de esto o, por decirlo así, el período de seguimiento en cada una de las dos intervenciones es corto.

En el ejemplo que venimos utilizando del β-HAT, sería totalmente inapropiado utilizar este tipo de diseño.

Debido a que debemos tener la certeza de que el efecto de la primera intervención no perdura hasta la segunda (ni la afecta), se ha desaconsejado en general. Sólo cuando hay una certeza prácticamente absoluta de este hecho sería utilizable este diseño.

ENSAYOS CLÍNICOS FRENTE A ENSAYOS COMUNITARIOS

Como se indicó al principio de este artículo, se incluye una breve consideración relativa a los ensayos comunitarios. Existen conceptos diversos relativos a qué es un ensayo clínico comunitario, pero el más generalizado es el que se discute a continuación.

Este diseño hace referencia a un diseño experimental en el que la unidad de intervención y de evaluación de la respuesta no es el individuo sino una comunidad.

En el diseño tradicional, se mide la variable de interés a nivel poblacional (p. ej., prevalencia de hipertensión arterial no controlada) en una comunidad (barrio, pueblo, ciudad, etc.), se realiza una intervención en toda la comunidad elegida (utilizando, por ejemplo, los medios de comunicación) y se mide, de nuevo a nivel poblacional, el efecto en la variable de interés. Paralelamente, existe una comunidad «control» en la que se realizan los mismos procedimientos salvo la intervención diseñada. Conceptos como aleatorización, doble ciego y otros pierden sentido cuando sólo hay una comunidad en la que se interviene y una comunidad control. Sin embargo, se han realizado ensayos comunitarios en los que múltiples comunidades (empresas, colegios, etc.) se han asignado aleatoriamente a intervención o control[18].

Aunque están alejados generalmente de la práctica habitual del cardiólogo clínico, el ensayo comunitario merece este breve comentario porque se ha utilizado a gran escala en cardiología en varias ocasiones, siendo la experiencia más notable la realizada en Karelia del Norte (Finlandia)[19].

Ya para terminar este artículo, se abordan de manera práctica dos preguntas con las que puede enfrentarse el cardiólogo que realiza o lee ensayos clínicos: ¿cuántos pacientes hay que estudiar?, y ¿cuándo debe pararse un ensayo clínico?

¿CUÁNTOS PACIENTES HAY QUE ESTUDIAR?

Esta es una pregunta con la que habitualmente se enfrenta cualquier investigador a la hora de realizar un estudio clínico del tipo que sea. En el caso de los ensayos clínicos, al leerlos puede surgir una pregunta similar formulada de otra manera: ¿son suficientes los pacientes que han incluido para la pregunta que quieren responder? o ¿qué capacidad tiene este estudio de detectar la diferencia que interesa detectar? La respuesta a esta última pregunta, si se pone en términos de probabilidad, es el *poder* del estudio. No se entra en discusiones adicionales de éste y otros conceptos que se recogen a continuación (test de una y dos colas o error α) porque ya han sido abordados en artículos previos de esta serie.

Como se apuntó en la introducción, no se van a realizar ni ofrecer en este artículo cálculos relativos al tamaño muestral, sino una reflexión de qué aspectos tener en cuenta al iniciar un estudio que nos orientarán al número de participantes necesarios en cada grupo. Es aconsejable entender estas reflexiones, necesarias antes del inicio de cualquier estudio, como *consideraciones del tamaño de la muestra* que nos conducen a un tamaño muestral necesario aproximado, pues están basadas en una serie de asunciones que pueden revelarse no exactas con el desarrollo del trabajo.

Una enumeración de los factores que influyen en el número necesario de participaciones es la siguiente:

a. Tasa esperada de eventos en ausencia de tratamiento (en el caso de que la evaluación de la respuesta sea de este tipo y no de modificación de una variable numérica como reducción de presión arterial o colesterol sérico).
b. Magnitud de la diferencia que se desea detectar.
c. Error α. Habitualmente 0,05 o menor.
d. Error β o poder del estudio (1 - error β).
e. Test de una o dos colas.
f. Pérdidas de seguimiento, abandonos.
g. Cambios de grupo terapéutico.
h. Factor efecto tiempo, cuando el efecto de la intervención no es de esperar que sea inmediato sino que es de esperar que empiece a observarse al cabo de cierto tiempo.

En el ejemplo que estamos utilizando, β-HAT, se estimó un tamaño muestral necesario de 4.020 pacientes. Se pensó que algunos centros tendrían problemas de reclutamiento de pacientes, por lo que se ajustó al alza esta cifra hasta 4.200. Las asunciones en las que se basó la estimación de 4.020 pacientes necesarios fueron que la mortalidad a los tres años en el grupo placebo sería del 18% y que el propranolol lograría, sin tiempo de demora, una reducción del 28% de la mortalidad en el grupo tratado; por otra parte, a lo largo de tres años, el 26% de los pacientes en el grupo de propranolol dejarían de tomar su medicación y al 21% de los pacientes en el grupo placebo se les prescribiría un beta-bloqueante. La mortalidad sería el evento principal de interés y se emplearía, para un test de dos colas, un error α de 0,05 y un poder (1-β) de 0,90. Al final del período de selección se habían incluido 3.837 pacientes, afectándose muy poco el poder del estudio por esta reducción del número de participantes[5].

Teniendo en cuenta los puntos anteriormente indicados, se está en disposición de calcular el número de pacientes necesario para un ensayo clínico y el poder de un determinado ensayo ya realizado para detectar diferencias, en caso de que los autores, hecho frecuente, no den indicaciones del mismo. A menos que uno tenga un particular interés por las fórmulas, es inútil intentar aprender la necesaria para ello pues hay programas sencillos de ordenador que ofrecen la respuesta. Es útil, sin embargo, utilizando uno de estos programas, realizar el ejercicio de explorar en qué medida la modificación de los factores sobre los que el investigador tiene capacidad de decisión (diferencia que se desea detectar y errores α y β) afecta al número de participantes necesarios.

¿CUÁNDO HAY QUE PARAR UN ENSAYO?

En una aproximación intuitiva, cuando se dé alguna de las siguientes situaciones: *a)* cuando el tratamiento

TABLA 4
Efecto sobre la significación estadística de los análisis repetidos de datos acumulados

N.º de tests repetidos con $\alpha = 0,05$	Nivel de significación global
1	0,05
2	0,08
3	0,11
4	0,13
5	0,14
10	0,19
20	0,25
50	0,32
100	0,37
1.000	0,53
∞	1,0

Adaptado de referencia 16.

TABLA 5
Significación nominal requerida para tests repetidos de dos colas en función del número de tests (N) y $\alpha = 0,05$

N	$\alpha = 0,05$
2	0,029
3	0,022
4	0,018
5	0,016
10	0,0106
15	0,0086
20	0,0075

Adaptado de referencia 16.

A es mejor que el B; *b)* cuando el tratamiento B es mejor que el A, y *c)* cuando es muy improbable que se detecte diferencia alguna. Esto, que parece obvio, se encuentra con algunos problemas logísticos. Uno de ellos es que si el tratamiento es mejor que el control (o viceversa) es deseable tener certeza de ello y en cuanto la tengamos proceder a parar el ensayo. Para ello, en particular en los ensayos que requieren largos seguimientos, tenemos que analizar los datos en más de una ocasión, con lo que, al revisar repetidas veces los mismos datos en diferentes fases, nos encontramos con que aumentamos la probabilidad de encontrar por azar una diferencia entre los grupos (error alfa). Es un problema de la misma índole que el que nos encontramos en las comparaciones múltiples de medias. Para permitir este análisis múltiple de datos se han propuesto múltiples métodos que, en resumen, adaptan el valor de p necesario para parar el ensayo en función del número de análisis que se planea realizar[8]. La tabla 4 presenta cómo afectan a la significación global del estudio los análisis repetidos de datos para dos tratamientos con respuesta normal, varianza conocida y separación idéntica entre análisis, aunque también es, a grandes rasgos, aplicable para otro tipo de datos[20]. En la tabla 5 se expone el valor facial de p requerido para disponer de un error α real de 0,05 dependiendo de el número de análisis repetidos de datos acumulados que se realicen. Al igual que en el caso anterior, están calculados para una variable con respuesta normal y varianza conocida pero son una buena aproximación en una gran variedad de situaciones[20].

Es importante señalar que en el proceso de esta toma de decisión la valoración estadística no es la única consideración que se tiene en cuenta. Otros ingredientes de este proceso son el impacto del momento de finalización del estudio en la interpretación del mismo, la naturaleza de preguntas no respondidas correspondientes a períodos más largos de tratamiento, así como la factibilidad de estudios futuros.

El β-HAT se paró nueve meses antes de lo planeado[4]. La figura 3 indica cómo se marcó la frontera para parar el estudio. Se planearon siete análisis intermedios y se utilizó el método de O'Brien para adaptar el valor de p en cada análisis, que, como se observa en la figura, hace más difícil una terminación precoz del ensayo y va acercando el valor de p en cada caso al deseado para el conjunto del estudio (que aproximadamente se alcanza en el momento del análisis final planeado)[21]. La línea de puntos indica el valor de Z = 1,96 o error $\alpha = 0,05$, que es el que se marcó en el diseño del ensayo. Vemos que a partir de la segunda revisión de los datos las diferencias ya excedían el nivel de p convencional (sin ajuste por la multiplicidad de análisis). En la sexta reunión del comité de monitorización, cuando el valor de p había excedido la frontera planeada con el método usado, se decidió para el ensayo[4].

GENERABILIDAD

Hasta ahora se han discutido, con mayor o menor profundidad, aspectos de validez interna de los ensayos clínicos. Sin embargo, existe un hecho al que no podemos escapar y es que los ensayos clínicos, y para el caso cualquier estudio, se realizan para aplicar sus conclusiones a otros pacientes externos al mismo. La pregunta de si los resultados de un ensayo clínico en particular son aplicables a la población de enfermos a la que atiende el clínico es una con la que se debe enfrentar a diario y para la que hay poco más que el sentido común para ayudarle en la decisión. Si bien el hecho de disponer de sólo una muestra (de las infinitas posibles) de la que queremos extender los resultados a una población más grande de la que procede es común a todos los diseños básicos en epidemiología (y única justificación de muchos de los tests estadísticos que se realizan), los ensayos clínicos añaden un problema adicional para su generabilidad y es el hecho de que,

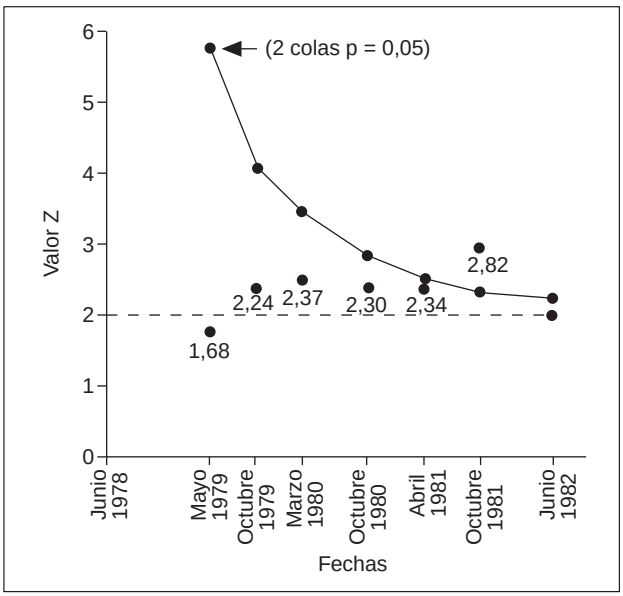

Fig. 3. Significación estadística de la diferencia de interés en las reuniones del comité de monitorización del β-HAT y frontera para parar el ensayo según método de O'Brien-Fleming[21].

generalmente, los pacientes participantes en un ensayo clínico son una muestra muy seleccionada como lo demuestra la menor mortalidad del grupo placebo en comparación con población general similar que se ha observado en muchos ensayos[17,22]. Cuanto más lejanos sean los pacientes incluidos en el ensayo a la población que atiende el clínico (en cuanto a características) más difícil es dar el necesario salto mortal de la generabilidad de los resultados. A este respecto, habitualmente es necesario realizar un juicio acerca de la aplicabilidad de los resultados a grupos no representados en absoluto en el estudio. Estos grupos pueden ser personas más jóvenes o mayores, mujeres en lugar de varones u otros grupos especiales. El juicio que hay que realizar se refiere a si biológicamente está justificado generalizar o si, por el contrario, debe esperarse a repeticiones del estudio en tales grupos. Esta última opción tiene unas implicaciones serias respecto a factibilidad, coste y posible demora en aplicar los beneficios de un tratamiento. A pesar de estos problemas de validez externa, los ensayos clínicos permanecen como la evidencia más sólida disponible para la realización de recomendaciones terapéuticas. Incluso se ha propuesto que la presencia de determinadas características en el diseño le confieran un determinado grado de solidez a las recomendaciones generadas de estos ensayos[23].

En resumen, a lo largo de este trabajo se han presentado y discutido algunos aspectos de interés en la lectura y desarrollo de ensayos clínicos. Este trabajo, idealmente, puede servir como una orientación para reflexionar sobre la calidad de la información aportada por un determinado diseño experimental, considerando cada caso individualmente y teniendo presente que la decisión final, muy probablemente, procederá de un conjunto de estudios y no de un estudio aislado.

BIBLIOGRAFÍA

1. Real Academia Española. Diccionario de la Lengua Española (20.ª ed.). Madrid: Espasa-Calpe, 1984.
2. Gabriel Sánchez R, Cabello López JB. Métodos de investigación en cardiología clínica. Introducción. Rev Esp Cardiol 1996; 49: 835-836.
3. Fernández-Avilés F. La metodología de la investigación en cardiología: una actualización necesaria. Rev Esp Cardiol 1996; 49: 834.
4. β-Blockers Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I.- Mortality results. JAMA 1982; 247: 1.707-1.714.
5. Byington RP for the Beta-Blocker Heart Attack Trial Research Group. Beta-Blocker Heart Attack Trials: Design, methods and baseline results. Control Clin Trials 1984; 5: 382-437.
6. Bradford Hill A. The clinical trial. Br Med Bull 1951; 7: 278-282.
7. Bulpitt CJ. Randomised controlled clinical trials. La Haya: Martinus Nijhoff Publishers, 1983.
8. Friedman LM, Furberg CD, De Mets DL. Fundamental of clinical trials (2.ª ed.). Littleton: PSG Publishing Company, Inc., 1985.
9. Snedecor GW, Cochran WG. Statistical methods (7.ª ed.). Ames: The Iowa State University Press, 1980.
10. Veterans Administration Cooperative Study Group on Antihypertensive Agents. Effect of treatment on morbidity in hypertension. Results in patients with diastolic blood pressure averaging 115 through 129 mmHg. JAMA 1967; 202: 1.028-1.034.
11. Veterans Administration Cooperative Study Group on Antihypertensive Agents. Effect of treatment on morbidity in hypertension. Results in patients with diastolic blood pressure averaging 90 through 114 mmHg. JAMA 1970; 213: 1.143-1.152.
12. Zelen M. A new design for randomized clinical trials. N Engl J Med 1979; 300: 1.242-1.245.
13. Angell M. Patient's preferences in randomized clinical trials. N Engl J Med 1984; 310: 1.385-1.387.
14. Ellenberg SS. Randomization designs in comparative clinical trials. N Engl J Med 1984; 310: 1.404-1.408.
15. Zelen M. Alternatives to classic randomized trials. Surg Clin North Am 1981; 61: 1.425-1.432.
16. Goy JJ, Eeckhout E, Burnand B, Vogt P, Stauffer JC, Hurni M et al. Coronary angioplasty versus left internal mammary artery grafting for isolated proximal left anterior descending artery stenosis. Lancet 1994; 343: 1.449-1.453.
17. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. N Engl J Med 1989; 321: 129-135.
18. Rose G, Tunstall-Pedoe HD, Heller RF. UK heart disease prevention project: incidence and mortality results. Lancet 1983; 1: 1.062-1.066.
19. Puska P, Toumilehto J, Salonen J, Nissinen A, Virtamo J, Björkqvist S et al. Community control of cardiovascular diseases. The North Karelia Project. Copenhague: World Health Organization, Regional Office for Europe, 1981.
20. Pocock SJ. Clinical trials. A practical approach. Nueva York: John Wiley & Sons, 1985.
21. De Mets DL, Hard R, Friedman LM, Lan KKG. Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. Controll Clin Trials 1984; 5: 362-372.
22. Castro Beiras A, Muñiz J, Juane R, Hervada J. Aspirina y corazón: Una llamada a la reflexión. Rev Esp Cardiol 1988; 41: 387-389.
23. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1992; 102 (Supl): 305-311.