Diseño estadístico de ensayos clínicos

Carlos Rubio Terrés

Laboratorios Lederle. División de Cyanamid Ibérica, S.A. Madrid.

investigación biomédica/ ensayos clínicos/ estadística/ tamaño muestral

El método en la investigación clínica

René Descartes (1596-1650) en su obra *Reglas para la dirección de la mente* estableció las directrices básicas del método científico. La regla IV indica que «para la investigación de la verdad de las cosas es necesario el método». La regla XIII propone que «si nosotros comprendemos perfectamente una cuestión, es preciso abstraerla de todo concepto superfluo, reducirla a su mayor simplicidad y dividirla en partes tan menudas como sea posible, enumerándolas»[1]. Estas reglas básicas, ideadas en el siglo XVII, siguen siendo válidas hoy día para el diseño de cualquier investigación y, por supuesto, para la investigación clínica.

La utilidad clínica de un tratamiento o medicamento se establece mediante dos tipos de razonamiento: el deductivo y el inductivo[2]. El razonamiento deductivo es el que va de lo general a lo específico. Es el que se plantea habitualmente el médico en la práctica clínica. Sabemos, por ejemplo, que un hipnótico determinado es un tratamiento eficaz del insomnio porque así ha sido demostrado en ensayos clínicos. Sin embargo, podemos preguntarnos, ¿cuál será su grado de eficacia en este paciente concreto? El razonamiento inductivo, por el contrario, va de lo específico a lo general. Supongamos que una compañía farmacéutica ha descubierto una molécula que tiene actividad hipnótico-sedante en animales y en voluntarios sanos. Pero, ¿es eficaz en el tratamiento del insomnio?, ¿es segura su utilización en dicha indicación?, ¿qué dosis son las adecuadas y en qué casos? La realización de estudios que respondan a estas y a otras cuestiones similares es lo que se denomina investigación clínica. El método más utilizado es el ensayo clínico, es decir, aquel estudio experimental y prospectivo en el cual el investigador provoca y controla las variables, y los pacientes son asignados de forma aleatoria a los distintos tratamientos que se comparan. Pero el ensayo clínico, entendido como el método científico así definido, es muy reciente. El primer ensayo clínico controlado y aleatorizado no se realizó hasta después de la Segunda Guerra Mundial, cuando Bradford Hill en 1946 estudió la eficacia de la estreptomicina en la tuberculosis pulmonar[3,4]. En la historia de la medicina son abundantes los ejemplos de tratamientos establecidos sin el debido rigor científico o ético. Recientemente se ha sabido que en la década de 1940, en el Hospital de la Universidad de California (EE.UU.) se inyectó plutonio a numerosos pacientes supuestamente terminales, sin su conocimiento ni consentimiento, para obtener información acerca de sus efectos tóxicos, con fines militares[5].

Otro conocido ejemplo es el del tratamiento de la esquizofrenia mediante el coma insulínico, propuesto por Manfred Sakel en 1933, sin suficiente base científica y desde luego sin realizar previamente estudios clínicos que demostraran la utilidad de dicho tratamiento[6]. A pesar de ello, fue de uso generalizado y aceptado hasta los años cincuenta, cuando la descripción de muertes y lesiones cerebrales en los pacientes tratados provocó la realización de un ensayo clínico cuyos resultados se publicaron en 1957. En el mismo no se demostró que el coma insulínico fuera superior al coma barbitúrico (que se supuso sin efecto) en la curación de la esquizofrenia[7]. Como consecuencia de dicho estudio, el tratamiento de la esquizofrenia mediante el coma insulínico fue progresivamente abandonado, con lo que se suprimió una práctica acientífica y probablemente dañina.

Queda clara, pues, la necesidad de realizar ensayos clínicos científica y éticamente válidos antes de instaurar un nuevo tratamiento. En la figura 1 se esquematiza la metódica general que debe seguirse en la realización de un ensayo clínico.

Definición del objetivo

Es el primer paso. Se trata simplemente de plantear una cuestión concreta, la hipótesis que se desea estudiar. Para ello debe realizarse una prueba de hipótesis. Si queremos comparar la efectividad de dos fármacos A y B, plantearemos una *hipótesis nula* (H_0), es decir A $\neq$ B (no habrá diferencias de eficacia entre ambos) y una *hipótesis alternativa*

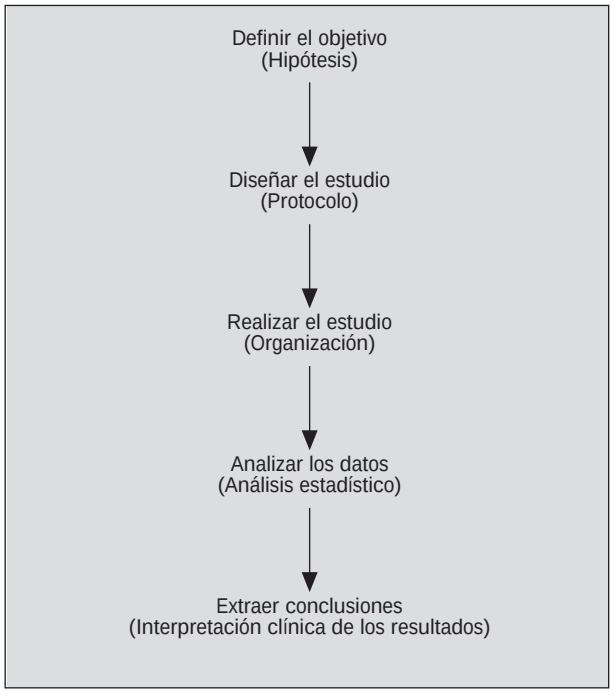

Fig. 1. El método en la investigación clínica.

Correspondencia: Dr. C. Rubio Terrés.
Virgen de Aránzazu, 21. 28034 Madrid.

Manuscrito aceptado el 19-12-1994

Med Clin (Barc) 1996; 107: 303-309

(H_1): A = B, por tanto A será mejor o peor que B. De acuerdo con el resultado del estudio, aceptaremos o, por el contrario, rechazaremos la hipótesis nula[8]. Una hipótesis puede calificarse como «buena» cuando es «posible» en lo que se refiere al número de pacientes que deben seleccionarse, al tiempo de realización del estudio, y al costo y medios que se precisan para poder responderla. Así mismo, debe ser «relevante» porque se entiende que aportará nuevos conocimientos o redundará en mejoras de la práctica clínica, y debe tener suficiente «fundamento», existiendo datos previos que la justifiquen. Por supuesto, debe ser de interés para el investigador y «ética» en su planteamiento[9].

Necesidad de grupo control

Es evidente que, en términos generales, para averiguar si un medicamento A es eficaz en una patología determinada, lo mejor es comparar inicialmente con un tratamiento no activo *(placebo).* Una vez demostrado que A es mejor que placebo, para determinar en qué medida es eficaz, lo lógico es comparar con un medicamento B *activo* y de eficacia ya conocida en esa patología. La incorporación de un grupo control tratado con placebo es importante también para establecer la comparación relativa de eficacia entre ambos tratamientos activos[10].

La comparación frente a un grupo *no tratado* sólo está justificada cuando no es posible la administración de un placebo por motivos éticos (p. ej., su administración continuada por vía parenteral, o debido a la gravedad de la patología, como en una septicemia) o bien porque es posible la valoración objetiva de los resultados (p. ej., mediante hemocultivo).

Además del tratamiento activo, del placebo y del no tratamiento, pueden realizarse estudios controlados comparando con *otras dosis* (estudios de búsqueda de dosis) o bien con *controles históricos*[10].

En estos últimos estudios surge la duda acerca de si la distribución de los factores pronósticos en el grupo histórico es similar o no a la del grupo contemporáneo. Para que los controles históricos puedan considerarse comparables a los controles contemporáneos, los pacientes deben proceder de los mismos hospitales y los métodos de evaluación de la respuesta, así como los criterios de inclusión/exclusión deben ser idénticos en ambos grupos. Otro aspecto importante a tener en cuenta es que la enfermedad debe tener un desenlace previsible, sin posibilidad de mejoría o curación espontánea que podría diferir en el tiempo y llevar a conclusiones erróneas. No obstante, debe tenerse en cuenta que a pesar de la adopción de medidas como las citadas, los estudios con controles históricos tienden a exagerar el valor de un nuevo tratamiento[11], por lo que sólo deben efectuarse por imposibilidad de establecer controles prospectivos, por ejemplo en enfermedades de muy baja incidencia, o cuando se prevean problemas serios en la selección de pacientes.

Los *autocontroles* (que cada paciente sea su propio control) pueden ser de utilidad para conocer la variabilidad intraindividual en determinados estudios, como en los de diseño cruzado de búsqueda de dosis. Sin embargo, para la evaluación de eficacia es preferible comparar distintos grupos de tratamiento.

Como principio, la realización de *estudios no controlados* debe ser la excepción. No obstante, éstos están justificados en determinados casos. No sería ético hacer estudios frente a placebo en pacientes con cáncer y privarles del tratamiento activo. No son controlados los estudios de farmacocinética en fase I. Tampoco algunos estudios iniciales en fase II. Respecto a los estudios de seguridad en fase IV, es muy conveniente que sean controlados ya que el perfil de seguridad de un medicamento sólo puede establecerse en comparación con placebo (que también puede inducir efectos adversos) o con un fármaco de referencia[12].

Significación y precisión estadísticas. Relevancia clínica

Supongamos que una vez terminado el ensayo clínico en el que comparamos la eficacia de los medicamentos A y B, obtenemos una tasa de respuesta del 50% con A y del 45% con B. Realizada la correspondiente prueba estadística, resulta que A es mejor que B, con una $p < 0,05$ y un IC 95% = 2,5 al 9%. ¿Qué es lo que esto significa? El principal objetivo de las pruebas estadísticas es responder a la pregunta ¿cuál es la probabilidad de que la diferencia observada se deba al azar?[2,13]. Una $p < 0,05$ indica que en menos de cinco veces de cada 100 que repitiéramos el mismo estudio, nuestro resultado se debería al azar. Una p significativa indica que existe una gran diferencia de efectividad entre ambos fármacos, o bien que la muestra estudiada es tan grande que hemos conseguido detectar una diferencia real, pero pequeña[13]. Debe advertirse que la *significación estadística* a partir de $p < 0,05$ es un convencionalismo, siendo también de interés la información que se obtiene de los resultados con una p algo mayor que 0,05. Cuanto mayor sea la p, más fuerte será la evidencia a favor de la hipótesis nula[2].

A veces surge la duda de si debemos elegir una p con una o con 2 colas. La respuesta es simple: puede calcularse con una cola siempre y cuando consideremos que el medicamento A nunca podrá ser peor que el fármaco de referencia (p. ej., un placebo). Sin embargo, incluso en la comparación de un fármaco activo con placebo, cabe la posibilidad de que se produzca un resultado falso negativo (que no se hallen diferencias frente al placebo). Esto es más frecuente cuando la variable medida incluye criterios subjetivos, por ejemplo en la depresión psíquica, o cuando la muestra estudiada es pequeña[10]. Por tanto, generalmente es conveniente elegir una p con 2 colas (A puede ser mejor o peor que B). Esta decisión es importante, ya que una $p = 0,04$ con una cola es estadísticamente significativa, mientras que la equivalente con 2 colas ($p = 0,08$) no lo sería.

La diferencia encontrada entre dos tratamientos puede ser estadísticamente significativa, y sin embargo no tener *relevancia clínica.* La relevancia clínica es un juicio subjetivo y pragmático acerca de la importancia real para la práctica clínica de la diferencia hallada. Por ejemplo, si comparásemos la tolerancia gastrointestinal de 2 antiinflamatorios A y B, el A no produjera trastornos gástricos en 90 de cada 100 pacientes y el B fuera, así mismo, bien tolerado en 93 de cada 100 tratados, aunque la diferencia fuese estadísticamente significativa ($p < 0,05$) parece evidente que no tendría relevancia clínica[13]. La *precisión estadística* es la inversa de la variancia (a mayor precisión, menor variancia) que aumenta con el aumento del tamaño de la muestra. Una manera de medir la precisión es mediante el intervalo de confianza (IC). Cuanto más amplio es el IC menor es la precisión, hay menor confianza en que el resultado no se deba al azar y pueda inferirse a la población[13].

Errores en la investigación clínica

Un resultado estadísticamente significativo (la consabida p) no supone ninguna garantía de que sea un resultado válido o real, que pueda inferirse a la población general con confianza, si no hemos cuidado el control de los errores estadísticos que se producen en los ensayos clínicos.

La *muestra* de pacientes incluidos en un ensayo clínico (p. ej., muestra de pacientes con infecciones ginecológicas de los hospitales de la provincia de Madrid) proviene de una *población de muestreo* (p. ej., totalidad de pacientes con infecciones ginecológicas de los hospitales de la provincia) y ésta, a su vez, proviene de la *población objetivo* (p. ej., pacientes españolas con infecciones ginecológicas)[2]. Se pretende que los resultados obtenidos en la muestra de pacientes puedan ser inferidos (generalizados) a la población objetivo. Para ello, deben controlarse los errores aleatorios y sistemáticos (o sesgos). Debe aclararse que ningún estudio está libre de errores y, por tanto, las inferencias nunca son perfectamente válidas[9]. Sin embargo, el objetivo es controlar al máximo posible dichos errores, para conseguir que:

1. La muestra seleccionada sea representativa de la población. Para ello, deben controlarse los *errores aleatorios,* aquellos que dan lugar a resultados erróneos debidos al azar. Su descontrol puede producir la distorsión de la muestra seleccionada (que la proporción de sus factores pronósticos sea diferente a la de la población). Por ejemplo, suponiendo que la prevalencia de sida en la población fuese del 30%, en una muestra bien seleccionada de individuos hallaríamos 30 con sida. Si en su lugar, el número de individuos con sida fuera de 20 o 40, esta distorsión podría deberse a un error aleatorio[2,9].
Como consecuencia del control de los errores aleatorios, se produce el aumento de la *precisión* y de la *significación* estadísticas de una estimación. Pero, ¿cómo controlar este tipo de error?: mediante el *muestreo al azar* y *aumentando el tamaño de la muestra (n)*[2,9].
2. La asignación de los pacientes de la muestra a los grupos de tratamiento debe dar lugar a una distribución similar de los factores pronósticos.
Deben controlarse, por tanto, los *sesgos (errores sistemáticos),* aquellos que producen resultados erróneos por defectos en el diseño del estudio. El grado de ausencia de sesgos va a determinar lo que se denomina *la validez estadística* del estudio. La *validez interna,* es decir, el grado de validez del resultado para los propios pacientes del estudio, y la *validez externa* o grado de confianza son la extrapolación o inferencia de ese resultado a la población[14].

Control de sesgos: aleatorización y enmascaramiento

Existen varios tipos de sesgos. Quizás uno de los principales es el de la *selección de pacientes* en el muestreo y en la asignación a los grupos de tratamiento[14]. Para controlar los sesgos del *muestreo* deben definirse correctamente los *criterios de inclusión/exclusión de pacientes,* para controlar adecuadamente los factores pronósticos conocidos (p. ej., en un ensayo clínico de un antimicrobiano en infecciones ginecológicas, serían posibles criterios de inclusión: mujeres hospitalizadas, de edad a partir de 16 años, con infecciones ginecológicas confirmadas mediante un diagnóstico clínico o bacteriológico y originadas por microorganismos patógenos de sensibilidad conocida o probable a el/los antimicrobianos en estudio; serían criterios de exclusión: pacientes alérgicas a antimicrobianos similares, disfunción renal de cierta intensidad, trastornos hepáticos, haber tomado otro antimicrobiano en las 72 h previas al reclutamiento, etc.). Así mismo, debe hacerse el muestreo *al azar*, con lo que se consigue el control de los factores pronósticos desconocidos (en el ejemplo anterior, supongamos que pueda existir un sector de la población femenina con una determinada característica o factor pronóstico desconocido que determine los resultados del estudio en infecciones ginecológicas; mediante el muestreo al azar, ese factor tendería a estar en la muestra en igual proporción que en la población).

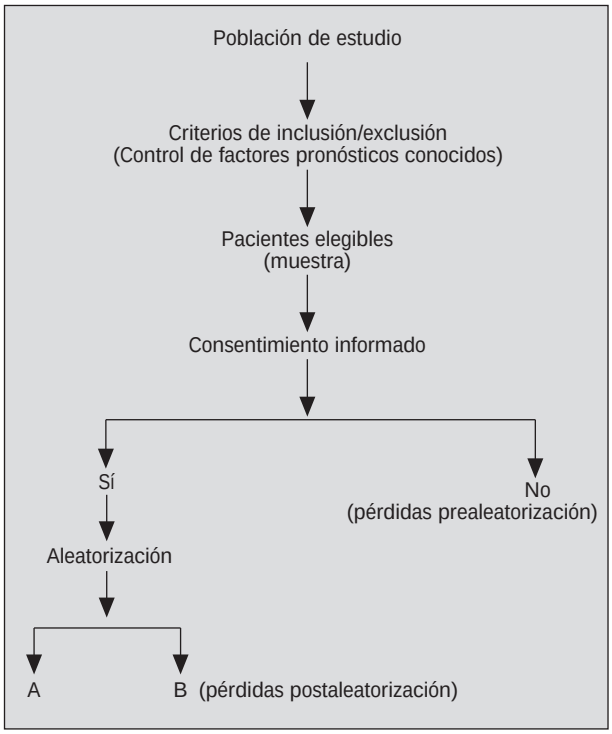

Fig. 2. Aleatorización: método general.

El otro sesgo en la selección es el que puede producirse en la asignación de los pacientes de la muestra a los grupos de tratamiento. Puede ocurrir que, de manera inconsciente o no, el investigador asigne a los pacientes con peor pronóstico al tratamiento ya conocido, en el que confía, y a los pacientes en mejor estado al tratamiento que se está probando. Para evitar este sesgo y, por tanto, con el objetivo de conseguir una distribución homogénea de los factores pronósticos, tanto de los conocidos como de los desconocidos, debe hacerse una *asignación aleatoria de los pacientes a los grupos de tratamiento (aleatorización).* El método general de la aleatorización se esquematiza en la figura 2. Con la asignación aleatoria no sólo se evitan los sesgos de asignación, sino que también se minimizan los propios errores aleatorios.
Los principales tipos y métodos de aleatorización (simple, por bloques, estratificada, centralizada, ciega y proporcional) son bien conocidos y están claramente descritos en la literatura[15,16]. Aunque debe valorarse caso por caso, en general puede decirse que la aleatorización por bloques es preferible a la simple, porque el número de pacientes por grupo es balanceado en la primera, y que, así mismo, es muy conveniente que sea ciega (la asignación de los tratamientos es desconocida para el investigador: estudios a doble ciego)[15]. Esto nos lleva al *sesgo de evaluación de los resultados* y a su control mediante el *enmascaramiento.* En un ensayo clínico no ciego (nadie –médico o paciente– desconoce la asignación de los tratamientos en estudio) es fácil que se introduzcan distorsiones (sesgos) en los efectos y valoraciones condicionadas –subjetivas y falsas– del tratamiento y de los resultados, tanto por el paciente como por parte del médico. Para evitarlo, se recurre al enmascaramiento, realizándose los denominados estudios a simple

ciego (tratamiento desconocido por el paciente), a doble ciego (también desconocido para el médico), a triple ciego (también para el monitor) y cuádruple ciego (también en el análisis estadístico)[15].
Los estudios no enmascarados pueden estar justificados en casos muy concretos: por imposibilidad galénica de realizar el enmascaramiento, por motivos éticos (p. ej., no es posible la administración parenteral continuada de un placebo), por las características del fármaco (p. ej., estudio de un antineoplásico del que deben hacerse ajustes de dosis).

El impacto de las pérdidas pre y postaleatorización

Como se mencionó anteriormente, los objetivos de la asignación aleatoria de los pacientes a los grupos de tratamiento son fundamentalmente dos: evitar los sesgos de asignación y minimizar los errores aleatorios.
Charlson et al[17] revisaron 41 ensayos clínicos aleatorizados existentes en la base de datos del Instituto Nacional de la Salud de los EE.UU. hasta 1979, hallando que las elevadas *pérdidas prealeatorización* (por la negativa de los pacientes a entrar en el estudio, o bien por ser considerados no elegibles de acuerdo con los criterios de inclusión/exclusión) (fig. 2) fueron responsables de que tan sólo el 34% de los estudios alcanzara la n prevista inicialmente. Estas pérdidas, además, disminuyen la validez externa de los estudios, ya que surge la duda acerca de si los pacientes no aleatorizados de la población de muestreo pueden ser diferentes en algún factor pronóstico respecto a los aleatorizados. Por ello, es muy importante hacer el seguimiento de los pacientes elegibles no alelatorizados[18].
¿Pero, de qué manera pueden reducirse estas pérdidas de pacientes antes de la aleatorización? En primer lugar, cuando sea posible, reduciendo el número de pacientes no elegibles, mediante criterios de inclusión y exclusión menos estrictos, es decir, realizando estudios más pragmáticos. En segundo lugar, consiguiendo que sea aleatorizado el mayor tanto por ciento posible de los pacientes elegibles (que cumplen los criterios de selección). Esto es fácil de decir, pero difícil en la práctica, ya que se trata de plantear al paciente la solicitud de su consentimiento de una manera «inteligente», para que, sin alarmarle pero al mismo tiempo informándole adecuadamente, se preste a participar en el ensayo clínico[17].
Las *pérdidas postaleatorización* (fig. 2) debidas a los abandonos y retiradas de pacientes durante el estudio son, si cabe, más graves. Destruyen la aleatorización y, por tanto, la validez interna y externa del ensayo. El tratamiento de estas pérdidas debe estar claramente especificado en el protocolo. Así mismo, deben tenerse en cuenta en el cálculo del tamaño de la muestra. Un método que evita la destrucción de la asignación aleatoria es el *análisis de los resultados por intención de tratar,* que incluye a todos los pacientes aleatorizados, hayan sido o no tratados. Este tipo de análisis es de gran importancia, ya que los resultados de los pacientes retirados pueden suministrar información muy valiosa. Un ejemplo es el del Helsinki Heart Study, acerca de la prevención primaria de la mortalidad coronaria mediante el hipolipemiante gemfibrozilo[19]. En el grupo tratado con el fármaco activo, se describió una tasa de mortalidad por accidentes, homicidios y suicidios algo superior a la observada en el grupo placebo[20], lo que ha llevado a plantear hipótesis acerca de los posibles efectos psiquiátricos derivados de las concentraciones bajas de colesterol.
La visión explicativa de los resultados es la que tiende a hacerse habitualmente, y consiste en el análisis sólo de los pacientes evaluables (no contempla los abandonos y las retiradas del estudio). La visión pragmática, como se ha dicho, es la del análisis por intención de tratar, de todos los pacientes aleatorizados. Ambos análisis (explicativo y pragmático) no son excluyentes y, por el contrario, puede ser de interés compararlos (tabla 1). La visión superpragmática consistiría en analizar a todos los pacientes elegibles (aleatorizados y no aleatorizados). Sin embargo, este tipo de análisis parece utópico, paradójicamente desde un punto de vista práctico, por lo que generalmente nos conformaremos con el seguimiento de las pérdidas prealeatorización, tal y como se explicó con anterioridad.

TABLA 1

Ensayos clínicos explicativos y pragmáticos*

Característica	Ensayo explicativo	Ensayo pragmático
Variable medida	Efecto del tratamiento en «condiciones de laboratorio»	Beneficio del tratamiento en la práctica clínica
Criterios de selección	Estrictos (pacientes homogéneos)	Laxos (todos los pacientes susceptibles de tratamiento)
Aleatorización	Sí	Sí
Enmascaramiento	Sí	No
Criterios de evaluación/objetivo	A > B (p < 0,05)	A es tratamiento de elección
Tipos de error	α y β	γ
Pacientes necesarios (n)	Para controlar α y β	Para controlar γ

*Tomados de referencia 31 (modificado).

El desequilibrio de factores pronósticos

A pesar de la aleatorización, por errores en la misma o bien debido al propio azar, puede ocurrir que al analizar los factores pronósticos (p. ej., edad, sexo, gravedad de la enfermedad, etc.) de los grupos de tratamiento, encontremos que existe un desequilibrio en alguno de ellos (p. ej., que haya más pacientes graves en un grupo que en otro, siendo la diferencia estadísticamente significativa). Este riesgo puede reducirse mediante la aleatorización estratificada. De ese modo, se asigna a los pacientes a unos subgrupos o estratos previamente establecidos, según factores pronósticos que puedan determinar los resultados del estudio[15]. Si en un estudio se considera que la edad es un factor pronóstico importante, puede estratificarse en pacientes de edad menor de 50 y de mayor o igual a 50 años, por ejemplo. En algún estudio puede ser de interés estratificar en pacientes ambulatorios y no ambulatorios[11], pacientes que han sufrido sólo un infarto o más de uno, etc. Qué factor, si es que hay alguno, tiene suficiente relevancia como para ser estratificado debe decidirse en cada ensayo clínico.
Es importante estratificar sólo un número reducido de factores (generalmente dos o tres como máximo) ya que la introducción de un número excesivo de variables a estratificar aumenta la complejidad del estudio y puede dificultar la consecución del número necesario de pacientes en cada estrato[15].
La evaluación del posible desequilibrio en los factores pronósticos es fundamental, lógicamente, en los estudios no aleatorizados, ya que en éstos el riesgo de desequilibrio es muy elevado.
Si se confirma el desequilibrio, debe realizarse el *ajuste de los factores pronósticos*, para comprobar si la diferencia entre los tratamientos sigue siendo significativa después del ajuste. El método de ajuste más utilizado para desequilibrios en uno o dos factores es el test de Mantel-Haenszel. En el

TABLA 2

Riesgos de resultados falsos en un ensayo clínico

Resultado en el ensayo clínico	Realidad		
	A > B	A = B	A < B
A > B	–	1/2 α	γ
A = B	β	–	β
A < B	γ	1/2 α	–

A > B: A mejor que B; A = B: A igual a B; A < B: A peor que B; α: falso positivo (riesgo α/error de tipo I); β: falso negativo (riesgo β/error de tipo II); γ: falso positivo pragmático (riesgo γ/error de tipo III).

TABLA 3

Tasa de respuesta a un tratamiento: intervalo de confianza y poder estadístico[24]

Tratamiento	Análisis intermedio	Análisis final
A	13/25 (52%)	500/1.000 (50%)
B	8/25 (32%)	450/1.000 (45%)
p	0,25	0,03
IC 95%	–7 al 47%	0,6 al 9%

IC 95%: intervalo de confianza del 95%.

caso de observarse diferencias significativas en tres o más de tres factores, para variables cuantitativas se aplica la regresión múltiple y para variables cualitativas la regresión logística[11].

Cuando no se hallan diferencias entre los resultados, analizados por intención de tratar, puede ser de interés realizar *análisis de subgrupos.* Estos análisis no son, en ningún caso, sustitutorios de la aleatorización estratificada y deben realizarse con prudencia, ya que las diferencias halladas en los subgrupos pueden llevar a conclusiones erróneas[12]. Un ejemplo es el del estudio The Multiple Risk Factor Intervention Trial (MRFIT), que no demostró un beneficio global originado por la intervención y, sin embargo, en el análisis de subgrupos se observó reducción de la mortalidad cardiovascular al cabo de un año en relación con la disminución del colesterol y el tabaquismo y, así mismo, en relación con la disminución de la tensión arterial. Estos resultados deben tomarse con gran prudencia[21,22].

El efecto del tamaño muestral

Una vez finalizado un estudio (A frente a B) podemos obtener tres resultados posibles: A es igual a B, o bien A es distinto de B (A es mejor o peor que B). Pero cualquiera de los resultados pudiera ser falso, es decir: el resultado del estudio puede no corresponder a la realidad. Pueden, por tanto, producirse resultados *falsos positivos* y *falsos negativos* (tabla 2).

Si encontramos una diferencia (por ejemplo, $p < 0,05$, con un IC del 95%) el *nivel de confianza* (1-α) será la probabilidad de que el IC contenga el valor real. La probabilidad de que la diferencia hallada y contenida en el IC no sea real es lo que se denomina *error de tipo I* o riesgo de *falso positivo*[23] α. El riesgo α se reduce y, por tanto, aumentan el nivel de confianza 1-α y la precisión del IC mediante el aumento del tamaño muestral (n)[23]. Consideración especial merecen los *ensayos clínicos negativos* ($p > 0,05$). Un resultado negativo (no se hallan diferencias) puede deberse a que realmente no exista diferencia entre los tratamientos o a que esta sea nimia y prácticamente indetectable. En ese caso, se trata de un verdadero negativo. Si, por el contrario, es un resultado *falso negativo* (no se detecta una diferencia que existe) puede deberse a dos motivos[23]: *a)* la existencia de sesgos en el estudio (en la selección de los pacientes, en la información recogida o en su análisis), o bien *b)* la falta de precisión estadística, debido a la existencia de un *error aleatorio de tipo II*, cuya probabilidad se denomina β (tabla 2). En este caso, se dice que el estudio carece del suficiente *poder estadístico* (1-β). El poder estadístico aumenta con el aumento de n.

¿Pero qué hacer ante un ensayo clínico negativo? En el ejemplo[24] de la tabla 3, observamos que se efectuó un análisis intermedio con tan sólo 25 pacientes tratados en cada grupo. La diferencia hallada no fue estadísticamente significativa y el IC 95% incluyó al 0, lo que confirma que no existía en ese momento una diferencia real entre A y B. Sin embargo, cuando se alcanzó la muestra deseada (n = 1.000 pacientes por grupo), entonces la diferencia sí fue significativa ($p = 0,03$), aunque el IC 95% estuvo cerca del 0 (0,6-9%) por lo que persisten algunas dudas acerca de la realidad de la diferencia hallada[24]. Vemos, así, la importancia de valorar la precisión (el intervalo de confianza) a la hora de analizar un resultado negativo (o positivo) y, así mismo, el efecto del tamaño muestral. Por tanto, ante un ensayo clínico negativo, las 2 preguntas que debemos hacernos son[23]: *1)* ¿qué información nos proporciona el intervalo de confianza?, y *2)* ¿es el número de pacientes estudiados suficiente para encontrar diferencias reales? En este caso, debe calcularse el poder estadístico a posteriori (1-β). Se considera que el poder del estudio es aceptable cuando está entre el 80 y el 90%.

En algunas ocasiones se plantea la realización de *estudios de equivalencia terapéutica,* es decir: ensayos clínicos que buscan demostrar igualdad entre los tratamientos. En primer lugar, debe decirse que intentar «probar la hipótesis nula (H_0)» es estadísticamente erróneo[23]. La H_0 que se plantea es, por ejemplo, que $A \geq B + d$, siendo A y B el tanto por ciento de éxitos en uno y otro grupo de tratamiento y «d» la diferencia máxima aceptable para confirmar la equivalencia. Si se rechaza la H_0, se considera que A = B porque no difieren más del valor «d». El objetivo de este tipo de estudios es conseguir demostrar equivalencia con una n pequeña, pero esto es, así mismo, erróneo, ya que la n necesaria para demostrar equivalencia es mayor que la precisa para demostrar diferencias (cuanto mayor es la diferencia entre dos tratamientos, menor es el número de pacientes que se necesita para poder demostrarlo)[23].

Puede ocurrir que en un análisis intermedio de un ensayo clínico, se observe que los tantos por ciento de respuesta hallados difieren de los inicialmente previstos y extraídos de la revisión de la bibliografía. En ese caso, podría estar justificado modificar el protocolo y prolongar el ensayo clínico, para alcanzar una muestra suficiente, capaz de detectar diferencias.

El *error de tipo III (riesgo γ)* podría denominarse como *falso positivo «pragmático»,* ya que es el que se produce cuando el resultado del ensayo indica que, por ejemplo, A es mejor que B, cuando la realidad es todo lo contrario: que A es peor que B (tabla 2). Esto es de suma importancia en un estudio pragmático que pretenda decidir cuál debe ser el tratamiento de elección (A o B). También se reduce el riesgo γ mediante el aumento del tamaño de la muestra (n) (tabla 2).

Hay numerosa bibliografía que se ocupa de revisar, con ejemplos, los métodos y fórmulas utilizados para el cálculo del tamaño muestral[3,11,25,26]. Aquí tan sólo comentaremos la fórmula general para el cálculo de n en la comparación de medias y porcentajes[3,25] y teniendo en cuenta los riesgos α y β (no el riesgo γ) válida, por lo tanto, para estudios expli-

cativos (no para estudios pragmáticos) (véase tabla 1, con las características de ambos tipos de estudios):

$$n = (2 \times V / \delta) \times f(\alpha, \beta)$$

donde *n* es el número total de individuos a estudiar en el ensayo. *V* es la variabilidad; si se trata de la diferencia de medias corresponde a $(S_1^2 + S_2^2)$; si se trata de la diferencia de porcentajes equivale a $(P_1 [100-P_1] + P_2 [100-P_2])$. δ es la diferencia (o nivel de sensibilidad) que se considera clínicamente relevante (p. ej., P_1-P_2 = una reducción del 5% en la mortalidad); este valor se evalúa y selecciona mediante la revisión de la bibliografía existente antes del inicio del estudio. *f* (α, β) es una constante que depende del nivel de protección frente a los errores de tipo I y II (se obtiene mediante tablas).

Análisis intermedios: ¿cuándo parar un EC?

La razón de ser de los análisis intermedios es que si, por ejemplo, en el estudio de dos tratamientos en una enfermedad grave (que produce mortalidad), se observan diferencias relevantes a favor de uno de ellos, por motivos éticos puede evitarse la prolongación innecesaria del estudio, consiguiendo que el grupo que recibe el tratamiento que parece menos favorable no se vea perjudicado. No obstante, este razonamiento es muy difícil de aplicar en la práctica[27]. Debe establecerse una, así llamada, *regla de interrupción,* consistente en fijar, en el protocolo del estudio, la variable que se analizará (p. ej., mortalidad), así como el número y momento de realización de análisis estadísticos intermedios que se llevarán a cabo a lo largo del estudio. Así mismo, se determinará el nivel de significación (o sensibilidad) que deberá alcanzar la diferencia entre los tratamientos, para efectuar la interrupción del estudio. Por ejemplo, de acuerdo con la regla de O'Brien[27] para mortalidad, deben realizarse 5 análisis en el plazo de un año, siendo los niveles de sensibilidad para decretarse la interrupción, los siguientes:

1.º: $p < 0{,}00000001$
2.º: $p < 0{,}0001$
3.º: $p < 0{,}001$
4.º: $p < 0{,}004$
5.º: $p < 0{,}009$

Existen otras reglas de interrupción, como la de Peto, consistente en 3 análisis[27], todos con una $p < 0{,}001$.
Un aspecto importante que afecta a las reglas de interrupción es el de la *corrección del exceso de pruebas estadísticas.* Aunque en la realidad A sea igual que B, el aumento en el número de pruebas estadísticas aumenta el riesgo de error α (de resultados falsos positivos)[11]. Por ejemplo, si repitiéramos 10 pruebas de significación al nivel del 5%, el nivel acumulado de significación, que sería de 0,05 en la primera prueba, al cabo de las 10 pruebas ascendería a 0,19 (fácilmente alcanzable, sin que existan diferencias reales)[11]. Por ello, si quisiéramos una p acumulada = 0,05 y estuviera previsto hacer 10 análisis intermedios, cada uno debería tener una regla de interrupción de $p < 0{,}01$[11].
En otro orden de cosas, debe comentarse brevemente la importancia de guardar la confidencialidad de los análisis intermedios. Deben ser ciegos para el investigador, con el propósito de evitar futuros sesgos en el caso de que el ensayo clínico prosiga. El investigador sólo debe conocer el resultado del análisis si se decide interrumpir el estudio. La mejor manera de guardar dicha confidencialidad es mediante la creación de un Comité independiente que se encargue del análisis. Así mismo, debe evitarse la publicación de los resultados mientras esté en marcha la aleatorización.

Un ejemplo reciente de la problemática de los análisis intermedios es el de la zidovudina en el tratamiento del sida[28]. Los primeros ensayos clínicos aleatorizados que se realizaron con este fármaco fueron interrumpidos porque se hallaron diferencias significativas frente al placebo en el recuento de CD4 por debajo de 500/µl. En 1989, de acuerdo con estos resultados y debido a la presión de los grupos activistas del sida en los EE.UU., la FDA decidió la aprobación de la zidovudina en los infectados por VIH asintomáticos[28]. Sin embargo, los resultados preliminares publicados con posterioridad, en 1993, correspondientes a un estudio europeo que no se interrumpió sino que prosiguió, el European Concorde Trial[29], no mostraron una diferencia clara a favor de la zidovudina en comparación con el placebo después de 3 años de seguimiento, ni en la supervivencia ni en la progresión de la enfermedad[29]. Este resultado hizo que en 1993 los activistas del sida de los EE.UU. pidieran a la FDA que, en lo sucesivo, retrasase la aprobación de nuevos agentes anti-VIH[28]. Este caso ilustra claramente el conflicto existente entre la responsabilidad de aleatorizar a un paciente individual a un tratamiento que parece ser peor inicialmente, y la responsabilidad o el compromiso de asegurarse del resultado, prosiguiendo el estudio, lo que redundará en el beneficio de miles o millones de futuros pacientes.

Conclusiones

La metodología estadística de la investigación clínica es motivo de preocupación y muestra de ello es la reciente publicación de un borrador de Guía al respecto del Comité de Especialidades Farmacéuticas (CEF o CPMP) de la Comisión de la Unión Europea[30]. Al comienzo de esta revisión, hablábamos del método científico y de su importancia en la investigación clínica. Es cierto que la metodología clínica y estadística y las buenas prácticas clínicas han complicado sobremanera la realización de ensayos clínicos. Sin embargo, debe tenerse siempre presente que el sujeto de la investigación es el ser humano y que, por tanto, los presupuestos éticos deben ir unidos a los metodológicos. Un ensayo clínico mal diseñado nunca debe hacerse ya que en ese caso se estaría incurriendo en el tan denostado «cobayismo». No obstante, la complejidad nunca deberá alcanzar cotas tales que impidan (por su elevado costo) el desarrollo de nuevos fármacos de interés sanitario. Sería, tal vez, útil profundizar en el papel de los denominados estudios pragmáticos[31] (tabla 1). En cualquier caso, y parafraseando a Séneca, deberemos convenir que «todo lo honesto es difícil»[32], y la investigación clínica no es una excepción.

REFERENCIAS BIBLIOGRÁFICAS

1. Descartes R. Discurso del método. Reglas para la dirección de la mente. Barcelona: Ediciones Orbis, S.A., 1983.
2. Plasencia A, Porta Serra M. La calidad de la información clínica (II): significación estadística. Med Clin (Barc) 1988; 90: 122-126.
3. Bakke OM, Carné X, García Alonso F. Ensayos clínicos con medicamentos. Fundamentos básicos, metodología y práctica. Barcelona: Doyma, S.A., 1994.
4. Streptomycin in Tuberculosis Trial Committee. Streptomycin treatment of pulmonary tuberculosis. Br Med J 1948; 2: 769-782.
5. Herken G, David J. Doctors in ardent pursuit of radiation wepons. International Herald Tribune 1994; 14 de enero: 6.
6. Spector R. Introduction to therapeutics. En: Spector R, editor. The scientific basis of clinical pharmacology. Principles and examples. Boston: Little, Brown and Co., 1986; 3-11.

7. Ackner B, Harris A, Oldham AJ. Insulin treatment of schizophrenia: a controlled trial. Lancet 1957; 1: 607.
8. González AG. Diseño y cálculo de tests estadísticos para ensayos clínicos y de laboratorio. Madrid: Escuela Universitaria de Enfermería, Universidad Complutense de Madrid, 1989.
9. Cummings SR, Browner WS, Hulley SB. Conceiving the research question. En: Hulley SB, Cummings SR, editores. Designing clinical research. Baltimore: Williams & Wilkins, 1988; 12-17.
10. Makuch RW, Johnson MF. Dilemmas in the use of active control groups in clinical research. IRB 1989; 11: 1-5.
11. Pocock SJ. Clinical trials. A practical approach. Chichester: John Wiley & Sons, 1983.
12. Circular 18/90. Directrices para la realización de estudios de Farmacovigilancia. Madrid: Dirección General de Farmacia y Productos Sanitarios, 21 de noviembre, 1990.
13. Porta Serra M, Plasencia A, Sanz F. La calidad de la información clínica (y III): ¿estadísticamente significativo o clínicamente importante? Med Clin (Barc) 1988; 90: 436-468.
14. Porta Serra M, Álvarez-Dardet C, Bolúmar F, Plasencia A, Velilla E. La calidad de la información clínica (I): validez. Med Clin (Barc) 1987; 89: 741-747.
15. Galende I, Tristán C. Problemas prácticos en un ensayo clínico (II). En: García Alonso F, Bakke OM, editores. Metodología del ensayo clínico. Barcelona: Fundación Dr. Antonio Esteve, 1991; 21-30.
16. Zelen M. The randomization and stratification of patients to clinical trials. J Chron Dis 1974; 27: 365-375.
17. Charlson ME, Horwitz R. Applying results of randomised trials to clinical practice: impact of losses before randomisation. Br Med J 1984; 289: 1.281-1.289.
18. De Abajo FJ, Serrano Castro MA. Problemas prácticos en un ensayo clínico (I). En: García Alonso F, Bakke OM, editores. Metodología del ensayo clínico. Barcelona: Fundación Dr. Antonio Esteve, 1991; 11-20.
19. Frick MH, Elo O, Haapa K, Heinonen OP, Heinsalmi P, Helo P et al. Helsinki Heart Study: primary-prevention trial with Gemfibrozil in middle-aged men with dyslipidemia. N Engl J Med 1987; 317: 1.237-1.245.
20. Frick MH, Heinonen OP, Huttunen JK, Manniven V. Gemfibrozil and coronary heart disease. N Engl J Med 1988; 318: 1.257.
21. Bulpitt CJ. Subgroup analysis. Lancet 1988; 2: 31-34.
22. Three analysis issues pertinent to designing an experiment. En: Hulley SB, Cummings SR, editores. Designing clinical research. Baltimore: Williams & Wilkins, 1988; 212-214.
23. Porta Serra M, Moreno V, Sanz F, Carné X, Velilla E. Una cuestión de poder. Med Clin (Barc) 1989; 92: 223-228.
24. Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986; 105: 429-435.
25. Carné X, Moreno V, Porta Serra M, Velilla E. El cálculo del número de pacientes necesario en la planificación de un estudio clínico. Med Clin (Barc) 1989; 92: 72-77.
26. Machin D, Campbell MJ. Statistical tables for the design of clinical trials. Oxford: Blackwell Scientific Publications, 1987.
27. Pocock SJ. When to stop a clinical trial. Br Med J 1992; 305: 235-240.
28. On stopping a trial before its time [editorial]. Lancet 1993; 342: 1.311-1.312.
29. Aboulker JP, Swart AM. Preliminary analysis of the Concorde Trial. Lancet 1993; 341: 889-890.
30. CPMP Working Party on Efficacy of medicinal products. Note for guidance: Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products (III/3630/92-EN, Draft 4). Bruselas: Commission of the European Communities, marzo 1993.
31. Vallvé C. Ensayos clínicos abiertos en investigación clínica. En: García Alonso F, Bakke OM, editores. Metodología del ensayo clínico. Barcelona: Fundación Dr. Antonio Esteve, 1991; 49-55.
32. Séneca LA. De los beneficios. En: Lucio Anneo Séneca: Obras completas. Madrid: Aguilar, 1961.