Cir Ciruj 2008;76:271-275

Ensayos clínicos sin significado estadístico. La importancia del error tipo II

Roberto Anaya-Prado, Fernando Grover-Páez, Ninel Mayari Centeno-López, Marisol Godínez-Rubí

Resumen

Un ensayo clínico aleatorio es un experimento prospectivo que compara una o más intervenciones contra un grupo control, con la finalidad de determinar la efectividad de las intervenciones. Un ensayo clínico puede comparar el efecto de una droga contra un placebo; o intervenciones médicas *versus* quirúrgicas. Los principios aplican en cualquier situación en la que el sujeto expuesto a un tratamiento está bajo el control del experimentador y el método de asignación es aleatorio. Un ensayo clínico negativo es aquel en el que no se encuentra una diferencia significativa entre los grupos comparados. Los resultados sin diferencia estadística pueden servir para desechar un tratamiento ineficaz o para demostrar que una intervención es tan efectiva como aquella con la que se le comparó. Eliminar tratamientos inútiles puede ser adecuado, pero si esto es el resultado de estudios con errores metodológicos, es posible que se prive a pacientes de nuevas intervenciones que en realidad son eficaces. En este artículo revisamos algunos posibles errores metodológicos que producen resultados falsos negativos en los ensayos clínicos.

Palabras clave: Ensayos clínicos, significancia estadística.

Summary

A randomized clinical trial is a prospective experiment to compare one or more interventions against a control group in order to determine the effectiveness of the interventions. A clinical trial may compare the value of a drug vs. placebo. It may compare surgical with medical interventions. The principles apply to any situation in which the issue of who is exposed to which condition is under the control of the experimenter, and that the method of assignment is through randomization. A negative clinical trial is that in which no significant difference is found between the comparison groups. Results without statistical difference may be useful either to discard useless treatments or to demonstrate that one intervention is as effective as the one it was compared with. Eliminating useless treatments may be adequate. However, if this is the result of studies with methodological errors, new interventions that are actually useful may not be available for patients. In this review we present some of the possible methodological errors that may lead to false negative results in clinical trials.

Key words: Clinical trials, statistical significance.

Introducción

La metodología estadística se usa para alcanzar conclusiones cuando se investiga qué tan compatibles fueron las observaciones de un estudio con la hipótesis de que el tratamiento experimental no tuvo ningún efecto (hipótesis nula). Cuando es poco probable que ocurran los resultados si esta hipótesis fuera verdadera, nosotros la rechazamos y concluimos que el tratamiento tuvo un efecto. Para ello utilizamos pruebas estadísticas (*F, t, q, q', z, o* χ^2), que cuantifican la diferencia entre las observaciones reales y las que esperaríamos si la hipótesis de "tratamiento sin efecto" fuera verdadera. Concluimos entonces que el tratamiento tuvo un efecto si el valor de esta prueba estadística es mayor a 95 % de los valores que ocurrirían si el tratamiento realmente no tuviera efecto. Cuando esto sucede, es común para los investigadores médicos reportar que su tratamiento en estudio tuvo un efecto "estadísticamente significativo". Pero cuando la prueba estadística no es suficientemente grande para rechazar la hipótesis de "tratamiento sin efecto", los investigadores registran "diferencia no estadísticamente significativa", y luego discuten sus resultados como si hubiesen demostrado que el tratamiento no tuvo efecto. Lo que realmente hicieron fue que fallaron en demostrar que el tratamiento tuvo un efecto. La diferencia entre demostrar en forma positiva que un tratamiento no tiene efecto, y fallar en demostrar que el tratamiento tiene un efecto, es sutil pero muy importante, especialmente considerando la pequeña cantidad de sujetos inclui-

Dirección de Educación e Investigación en Salud, Unidad Médica de Alta Especialidad, Hospital de Ginecoobstetricia, Centro Médico Nacional de Occidente, Instituto Mexicano del Seguro Social, Guadalajara, Jalisco.

Solicitud de sobretiros:
Roberto Anaya-Prado,
Blvd. Puerta de Hierro 5150, edificio B,
segundo piso, despacho 201-B,
Frac. Corporativo Zapopan,
45110 Zapopan, Jalisco, México.
Tel. y fax: (33) 3848 5410.
E-mail: robana@prodigy.net.mx

Recibido para publicación: 12-06-2007
Aceptado para publicación: 09-07-2007

dos en la mayoría de los ensayos clínicos.[1] Este problema es encontrado particularmente en ensayos clínicos pequeños en los que "no hay fallas" en el grupo de tratamiento. Esta situación nos puede llevar a resaltar evaluaciones exageradamente optimistas acerca de la eficacia terapéutica de una droga o una intervención quirúrgica.

La capacidad para identificar el efecto de un tratamiento con cierto nivel de confianza, depende del tamaño del efecto del tratamiento, de la variabilidad dentro de la población, y del tamaño de las muestras usadas. De la misma manera en que muestras grandes hacen más fácil detectar un efecto en el tratamiento utilizado, muestras más pequeñas lo hacen más difícil. En términos prácticos, esto significa que los estudios de terapias que involucran a unos cuantos sujetos y fallan en rechazar la hipótesis de "tratamiento sin efecto", pueden concluir este resultado porque los procedimientos estadísticos carecieron de *poder* para identificar el efecto debido a un tamaño de muestra muy pequeño, aun cuando el tratamiento sí tuvo un efecto. Por el contrario, las consideraciones acerca del poder de la prueba permitirán calcular el tamaño de la muestra necesaria para identificar un efecto de tratamiento de un tamaño dado que se cree está presente.[2]

Este artículo examina cómo estudios correctamente diseñados y que emplean métodos estadísticos en forma apropiada, pueden fallar en identificar diferencias (quizá clínicamente importantes) simplemente porque los tamaños de muestras son demasiado pequeños; además de otros posibles errores metodológicos que producen resultados falsos negativos en los ensayos clínicos.

Poder estadístico inadecuado

Un ensayo clínico debe tener un tamaño de muestra lo suficientemente grande para tener mayor probabilidad de encontrar una "diferencia verdadera" entre los grupos de estudio (poder estadístico). Si se realiza un ensayo pequeño y no se encuentra diferencia significativa, no se obtiene información nueva; no se encontró una diferencia simplemente porque no se tienen suficientes sujetos en el estudio. Por lo tanto, no se puede afirmar que no existe diferencia entre los tratamientos. Por el contrario, si tenemos un estudio grande y no se encuentra diferencia significativa, entonces podemos decir con mayor seguridad que los tratamientos realmente no son diferentes. El valor de p es un índice de poder de evidencia con relación al rechazo de la hipótesis nula. Hay quienes piensan que un valor de p implica el mismo peso sin importar si viene de un estudio grande o pequeño. Otros piensan que si se obtiene un resultado significativo en un ensayo clínico pequeño, esto quiere decir que el efecto (o la diferencia entre dos medias poblacionales) debe ser suficientemente grande para poder detectarlo incluso con muestras pequeñas, y, por lo tanto, es una diferencia significativa. Es verdad que si el tamaño de la muestra es suficientemente grande, podemos encontrar significado estadístico si la diferencia real entre las medias es muy pequeña y prácticamente irrelevante. Por lo tanto, encontrar una diferencia significativa en un ensayo clínico pequeño significa que el efecto fue relativamente grande.[3]

En 1979, Freiman y colaboradores examinaron 71 ensayos clínicos aleatorios publicados entre 1960 y 1977 en revistas como *Lancet, New England Journal of Medicine* y *Journal of the American Medical Association* (JAMA), que reportaban que el tratamiento estudiado no produjo una mejoría "estadísticamente significativa" ($p < 0.05$) en su resultado clínico final. Solamente 20 % de estos estudios incluía los sujetos suficientes para detectar 25 % de mejoría en el resultado clínico, con un poder de 0.50 o mejor.[4] En otras palabras, si el tratamiento produjo 25 % de reducción en la mortalidad u otra mejoría clínicamente importante, hubo al menos una posibilidad de 50:50 de que el ensayo clínico se pudiera haber detectado con una $p < 0.05$. Además, Freiman y colaboradores encontraron que solo uno de los 71 artículos indicó que fueron considerados los errores α o β al inicio del estudio; 18 reconocieron una tendencia en los resultados, en tanto que 14 comentaron la necesidad de un tamaño mayor de la muestra. Quince años más tarde, en 1994, Moher y colaboradores retomaron esta interrogante y analizaron los ensayos clínicos controlados aleatorios en estos mismos seriales publicados en los años de 1975, 1980, 1985 y 1990.[5] En tanto que el número de ensayos clínicos controlados aleatorios publicados en 1990 fue más del doble que los publicados en 1975, la proporción que reportaba resultados negativos permaneció razonablemente constante, aproximadamente 27 % de todos los ensayos clínicos. Solamente 16 y 36 % de los estudios negativos tuvieron un poder adecuado (0.80) para identificar 25 o 50 % de cambio en los resultados finales, respectivamente. Solamente un tercio de los estudios con resultados negativos indicaron información con relación a cómo se computó el tamaño de la muestra. Aunque un tanto desalentadores, estos resultados fueron un poco mejores que los reportados por Frieman y colaboradores en 1979, cuando nadie reportó cálculos de tamaño muestral.

Un poder estadístico inadecuado es quizá la causa más común de ensayos clínicos negativos. Para calcular el tamaño de una muestra para un estudio se tiene que especificar el efecto del tamaño, el nivel del significado estadístico y el poder deseado. Aunque el número de errores potenciales en medicina clínica es casi ilimitado, en la investigación clínica solamente pueden ocurrir dos errores básicos: encontrar una diferencia cuando realmente no existe (un error falso-positivo o error tipo I) y viceversa (un error falso-negativo o error tipo II). El error falso-positivo habitualmente es medido con la letra α, que usualmente se fija en 0.05 (el valor de p tradicional). Esto significa que los investigadores están dispuestos a aceptar una posibilidad de 1 en 20 de cometer este error. La conclusión falsa-negativa, por su lado, es medida con β, que habitualmente se fija más alto, por ejemplo 0.10 o 0.20. Esto significa que los investigadores están dispuestos a aceptar una posibilidad de 10 o 20 % de decir que no existe diferencia, cuando en realidad existe una deferencia en la población. El poder de un estudio es su capacidad para evitar un error

tipo II; matemáticamente, "poder" es igual a $1 - \beta$. Una descripción verbal sería que "poder" es la capacidad de un estudio para encontrar una diferencia (estadísticamente significativa con un valor de $p < \alpha$) si existe en la población. Por ejemplo, con un poder de 80 % y α de 0.05, un estudio tendría una posibilidad de 80 % de encontrar una diferencia estadísticamente significativa (en $p < 0.05$) si existe la diferencia en la población.

Se necesitan otros dos números para calcular tamaños de muestra con resultados dicotómicos (binarios), tales como enfermo o no enfermo. Éstas son estimaciones de la frecuencia de los resultados en los dos grupos. Una manera útil de abordar esto es estimar la frecuencia de un evento en la población y luego estimar qué tan grande esperaríamos que fuera el efecto del tratamiento. Así se necesitan cuatro números para calcular tamaños de muestras para ensayos con resultados dicotómicos: error α, error β, proporción del resultado en el grupo tratado, y proporción del resultado en el grupo control. Existen muchos programas que ayudan a calcular el tamaño de la muestra y su poder, tales como Epi-info.[6] Todos tienen el fuerte atractivo de evitar algunas operaciones matemáticas que se requieren para hacer el cálculo manualmente. A continuación, un ejemplo que ilustra el impacto del tamaño de la muestra:

Consideremos que se realiza un ensayo clínico con α de 0.05 y un poder de 0.90, con un resultado de 6 % en el grupo de tratamiento y 10 % en el grupo control. Esto refleja una reducción absoluta anticipada de 4 % en la frecuencia del resultado. Con estas presunciones se necesitarían aproximadamente 965 pacientes en cada grupo, para un total de 1930. Si se buscara un efecto protector más pequeño, digamos 8 % (reflejando solo una reducción absoluta de 2 %), se necesitaría un tamaño de muestra más grande: 4301 participantes en cada grupo, o bien, 8602 participantes en total. Por lo tanto, partir en dos el efecto absoluto del tratamiento (reducción de 4 % *versus* reducción de 2 % con tratamiento) cuadriplica el tamaño de la muestra. Si se buscara un efecto del tratamiento aún más pequeño, el efecto en el tamaño de la muestra sería todavía mayor. Al buscar una reducción del riesgo absoluto de solamente 1 % (10 % en el grupo control *versus* 9 % en el grupo de tratamiento), el tamaño de la muestra crecería más de 18 veces en cada grupo de tratamiento.

Un estudio no estadísticamente significativo (negativo) debe discutir la posibilidad de un tamaño inapropiado de la muestra como una causa probable del resultado no significativo, de lo contrario será la acuciosidad del lector la que determinará si se incluyeron suficientes pacientes en el estudio.[7]

Contaminación del estudio

Idealmente en un ensayo clínico controlado solo los pacientes asignados a la intervención experimental son los que la reciben. Sin embargo, algunas veces un buen número de pacientes asignados al grupo control reciben la intervención experimental, ya sea porque los médicos a cargo están convencidos de las bondades no demostradas de la terapia en estudio, o porque de alguna manera los pacientes se las ingenian para no estar en un grupo control e intercambian la mitad de los tratamientos entre ellos. Al mezclar las intervenciones se disminuye el poder estadístico del estudio y así la probabilidad de encontrar una diferencia real entre los grupos. La contaminación es más probable en ensayos clínicos donde el tratamiento no es fácil de cegar, y donde la aceptación o bondades del tratamiento experimental hacen que los médicos o los pacientes teman perder la oportunidad de los beneficios de la terapia en estudio.

Sobreutilización del análisis de intención de tratar

En general, los datos en los ensayos clínicos deben ser analizados comparando los grupos tal como fueron asignados mediante el método aleatorio, y no comparando con el grupo control (placebo) solo los pacientes en el grupo de tratamiento que sí recibieron la droga. La población asignada al grupo de la droga activa debe ser incluida con ese grupo para su análisis, aunque no haya recibido el medicamento. Esto puede sonar extraño: ¿cómo puede evaluarse la eficacia de una droga si el paciente no la toma? La razón real por la que la gente no cumple con el régimen puede tener que ver con los efectos adversos del medicamento, de tal manera que si se selecciona solo a los pacientes que cumplieron con el régimen, tendremos a un grupo diferente del de asignación aleatoria y podemos tener una imagen sesgada de los efectos de la droga. La inclusión en el análisis de los no cumplidores diluye el efecto, de tal suerte que se tiene que hacer todo lo posible para disminuir el número de estos "no cumplidores". En algunos ensayos clínicos, una buena medida para favorecer el apego al tratamiento son los criterios de inclusión, con esto se hace una evaluación de cada paciente para determinar sus posibilidades de adherencia al régimen. Los pacientes que tienen bajas probabilidades de cumplimiento se excluyen antes de la asignación aleatoria. Sin embargo, si la pregunta a la mano es qué tan aceptable es el tratamiento para el paciente, además de su eficacia, entonces la base para la inclusión puede ser la población en general que se beneficiaría de la droga, incluyendo a los no cumplidores.

Falta de apego al tratamiento

La falta de apego al tratamiento experimental disminuye la capacidad de un estudio para encontrar una diferencia entre dos tratamientos. Esto sucede cuando los pacientes no toman el tratamiento experimental o lo toman de manera irregular, particularmente cuando el régimen es complejo, cuando el tratamiento experimental produce efectos colaterales, y cuando existe poca motivación del paciente para seguir las instrucciones. En estas circunstancias, cuando el paciente es evaluado en el seguimiento puede referir no encontrar mejoría con el tratamiento proporcionado o "inventar"

que se siente mejor por temor a que descubran que no toma el tratamiento. Evidentemente, la interpretación del resultado será errónea cuando se analicen los resultados al abrir el cegamiento. En un estudio que no muestra diferencia entre los grupos tratados se espera encontrar una descripción de cómo se evaluó el apego al tratamiento, por ejemplo verificando los niveles de la droga o sus metabolitos en la orina o sangre del paciente, o la administración personalizada de la droga correspondiente por personal asignado.[8]

Seguimiento insuficiente o prolongado

Algunas veces el análisis de los resultados de un estudio se realiza prematuramente y no es posible identificar efectos de la intervención. Por el contrario, cuando el seguimiento es muy largo se disminuye la posibilidad de mantener una diferencia entre el efecto de los tratamientos. Un buen ejemplo es la comparación entre tratamientos médicos y quirúrgicos, un seguimiento breve desfavorece fuertemente a la cirugía por las complicaciones y efectos adversos a corto plazo (como dolor), cuando sus efectos verdaderos se aprecian a plazos más prolongados. Hacer seguimiento a un año, por ejemplo, en plastias inguinales, asegurando la no recurrencia es completamente inadecuado, toda vez que el seguimiento debe ser superior a los 3.7 años para poder asegurar que no existió verdaderamente recurrencia con la técnica empleada. Esto sería un seguimiento insuficiente.[9]

Competencia de riesgo

Muchas veces se ignora (especialmente en los ancianos) las comorbilidades de los pacientes en estudio. Es decir, tienen mayor riesgo de morir por causas ajenas a la enfermedad en estudio, limitando las probabilidades de encontrar una diferencia real entre los grupos tratados. Es decir, si se busca estudiar la disminución en la morbilidad o mortalidad con un medicamento específico y para una determinada enfermedad, el paciente puede morir por una causa ajena a la enfermedad que se busca tratar, y se puede interpretar negativamente que la droga en estudio no sirve para disminuir esa morbilidad o mortalidad. Las comorbilidades del paciente compiten y se limita la posibilidad de identificar el efecto real de la droga en estudio.

Clasificación errónea del resultado en estudio

Puede presentarse una información sesgada (o clasificación equívoca) cuando existe una imprecisión sistemática en la medición. Esto se puede visualizar mejor en los estudios epidemiológicos que involucran exposiciones dicotómicas y variables de enfermedades tales como niveles séricos elevados de colesterol e infarto del miocardio. Los sujetos se clasifican de acuerdo a la coincidencia de niveles séricos elevados de colesterol e infarto del miocardio. El investigador puede estar correcto o incorrecto, resultando en un hallazgo verdadero positivo o verdadero negativo, así como clasificaciones falsas positivas y falsas negativas de los sujetos respecto a la exposición y la enfermedad. El problema es mayor cuando el resultado de un estudio es susceptible a gran variabilidad intra o interobservador. Si la clasificación errónea sucede en un número considerable de pacientes y no es sistemática, se corre el riesgo de no detectar una diferencia significativa entre los grupos. En un estudio con resultados negativos se debe buscar información pertinente sobre la manera como se definieron y evaluaron los eventos a comparar. También se debe buscar la variabilidad inter e intraobservador al clasificar los eventos, con la finalidad de considerar que no fue la clasificación errónea de los resultados lo que produjo un ensayo clínico con resultados negativos. Como puede apreciarse, este problema depende básicamente de la interpretación subjetiva de resultados. De allí que se debe definir claramente en la metodología de un estudio la forma como se midieron las variables para arribar a la interpretación que finalmente será interpretada como estadísticamente significativa o no estadísticamente significativa.

Método inadecuado de asignación aleatoria

El principio básico en el diseño de ensayos clínicos o cualquier investigación científica es evitar errores sistemáticos. El conocimiento de las variables pueden afectar el resultado de un experimento; la mejor manera de evitar errores sistemáticos es asignando los individuos a los grupos en forma aleatoria. La asignación aleatoria tiene como finalidad asegurar una distribución igual o aproximada de las variables, entre los grupos de individuos en el estudio. Otro propósito del método de asignación aleatoria tiene que ver con el hecho de que las técnicas estadísticas utilizadas para comparar los resultados entre los grupos de pacientes en estudio son válidas bajo ciertas asunciones que surgen de la asignación aleatoria. Debe recordarse que algunas veces la asignación aleatoria no produce grupos comparables debido al azar. Esto puede representar un problema mayor en la interpretación de los resultados, dado que las diferencias en los resultados pueden reflejar diferencias en la composición de los grupos en las características basales más que en el efecto de la intervención. Existen métodos estadísticos para ajustar las características basales que se conoce están relacionadas al resultado. Algunos de estos métodos son la regresión logística, los modelos proporcionales de Cox y el análisis de regresión múltiple. En un estudio sin diferencia estadística se debe determinar si las características que pueden estar asociadas a un resultado adverso, se distribuyeron equitativamente entre los grupos.

A pesar de su importancia para evitar los sesgos, frecuentemente la asignación aleatoria no se realiza bien o no se efectúa

en los ensayos publicados. Algunas técnicas llamadas "aleatorias" no lo son; esto apareció en 5 % de una muestra de ensayos examinados.[10] Un ejemplo de técnica no aleatoria es asegurar que se realiza asignación aleatoria utilizando los días o semanas alternas para asignar tratamientos.[11] Los abordajes preferidos para llevar a cabo una secuencia de asignación aleatoria incluye las tablas de números aleatorios o los números aleatorios generados por una computadora.[12] Usar números nones o pares para asignar a los participantes a dos brazos de tratamiento es el abordaje más simple (asignación aleatoria simple). La asignación aleatoria restringida es un abordaje alternativo. Una forma común de la asignación aleatoria restringida es formando bloques. Este proceso asigna aleatoriamente a los participantes en una serie de bloques de tamaños especificados, de seis u ocho participantes. A cada uno de estos bloques se le da un número o rango de números, y la secuencia de seis participantes en el bloque es escogida con una tabla de números aleatorios o una secuencia de números generada por computadora. Así, después de cada sexto participante (por ejemplo 6, 12, 18, 24, etc.), habrán sido asignados números iguales a "A" y "B". Entre más grande sea el bloque, es menos probable que los investigadores inquisitivos y el *staff* en formación, tengan la capacidad de descifrar la longitud del bloque y así la siguiente asignación.[13]

Seguimiento sesgado

En un ensayo clínico o en un estudio de cohorte uno de los problemas potenciales lo representa la pérdida del seguimiento. Una vez que los sujetos están incluidos en un estudio, ellos pueden decidir en cualquier momento dejar de participar. Algunos individuos son más propensos a salirse del estudio que otros. También durante el curso del estudio algunos sujetos pueden morir por causas diferentes a la variable de interés. Aún más importante, algunos pacientes eligen no acudir más al estudio por mejoría sustancial de su condición clínica. Así, en un estudio negativo no es sencillo determinar si la pérdida de pacientes en el seguimiento fue sistemática, y los autores deben dejar claro que el seguimiento tuvo la misma intensidad en ambos grupos y que el tratamiento fue cegado. Es decir, las razones por las que se perdieron (excluyeron) los pacientes del seguimiento fueron similares entre los grupos comparados. De entrada, las pérdidas de seguimiento no parecen estar relacionadas a la selección dado que el sujeto ya estaba incluido en el estudio. Sin embargo, si los sujetos perdidos difieren en su riesgo de la variable de interés, se obtienen estimaciones de riesgo sesgadas. También si las manifestaciones tempranas no reconocidas de la enfermedad de interés hizo que las personas expuestas dejaran el estudio más frecuentemente que las no expuestas, se puede llegar a conclusiones distorsionadas.

Referencias

1. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? JAMA 1983;249:1743-1745.
2. López-Jiménez F, Paniagua D, Lamas GA. Interpretación de ensayos clínicos negativos o sin diferencia. En: López-Jiménez F, ed. Manual de medicina basada en evidencia. México: El Manual Moderno/JGH Editores;2001. pp. 99-109.
3. Greenberg RS, Daniels SR, Flanders WD, Eley JW, Boring JR. Interpretation of epidemiologic literature. In: Greenberg RS, ed. Medical Epidemiology. 3rd ed. New York: McGraw-Hill;2001. pp. 175-188.
4. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized controlled trial. N Engl J Med 1978;299:690-694.
5. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-124.
6. Dean AG, Dean JA, Burton AH, Dicker RC. Epi Info: a general-purpose microcomputer program for public health information systems. Am J Prev Med 1991;7:178-182.
7. Grimes DA, Schulz KF. Clinical research in obstetrics and gynecology: a Baedeker for busy clinicians. Obstet Gynecol Surv 2002;57:S35-S53.
8. Cuzick J, Edwards R, Segman N. Adjusting for non-compliance and contamination in randomized controlled trials. Stat Med 1997;16:1017-1029.
9. Howard G, Chambless LE, Kronmai RA. Assessing differences in clinical trials comparing surgical vs nonsurgical therapy. JAMA 1997;278:1432-1436.
10. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994;272:225-228.
11. Isemberg SJ, Apt L, Word M. A controlled trial of pivodine-iodine as prophylaxis against ophthalmia neonatorum. N Engl J Med 1995;332:562-566.
12. Schulz KF, Grimes DA. Generation of allocation sequences in randomized trials: chance, not choice. Lancet 2002;359:515-519.
13. Wassertheil-Smoller S. Mostly about clinical trials. In: Wassertheil-Smoller S, ed. Biostatistics and Epidemiology, 2nd ed. New York: Springer-Verlag;1990. pp. 129-146.