

Tamaño de muestra de protocolos clínicos

Alvar Loria

Secretaría de Salud, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Unidad de Epidemiología Clínica, Ciudad de México, México

Resumen

El cálculo del tamaño de muestra en los protocolos clínicos que se revisan en el comité institucional del Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán tiene básicamente tres tipos de problemas: datos insuficientes para recalcular el tamaño de la muestra, uso de ecuaciones obsoletas que no toman en cuenta el error beta y uso de una ecuación que no corresponde al diseño del estudio. En este documento discutimos los cuatro datos que deben ser incluidos en los protocolos de los estudios clínicos aleatorizados que comparan medicamentos versus placebo en los estudios de superioridad y versus control activo en los estudios de no inferioridad.

PALABRAS CLAVE: *Tamaño de muestra. Ética. Investigación clínica. Estudios de no inferioridad.*

Abstract

Sample size calculation in clinical protocols submitted to the review board of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán frequently has three types of problems: insufficient data for sample size recalculation, use of obsolete equations that do not take the beta error into account, and use of equations that do not correspond to the study design. In this document, we discuss the four data that should be included in the clinical protocols of randomized controlled trials that compare drugs versus placebo in superiority trials, and versus active controls in non-inferiority studies.

KEY WORDS: *Sample size. Ethics. Clinical research. Non-inferiority trials.*

El conocimiento no solo descansa en la verdad sino también en el error

CARL GUSTAV JUNG

¿Por qué es importante el cálculo de tamaño de muestra [TM]

Actualmente se considera una falta ética incluir un déficit o un exceso de sujetos de estudio en una investigación clínica. Debe ingresarse un número que permita detectar, con cierta certeza, el efecto de una intervención médica en caso de que la haya, y simultáneamente, cierta certeza de que si no hay efecto, la intervención sea verdaderamente negativa.

El exceso de sujetos de estudio aumenta las probabilidades de daño iatrogénico a los sujetos y, además, encarece los estudios. Su opuesto, una muestra insuficiente, de poco poder, puede llevar a estudios clínicos negativos, lo cual conduce a menor probabilidad de ser publicado y esto a su vez, al llamado sesgo de publicación que se traduce en un déficit de estudios negativos.¹

El sesgo de publicación lleva a aumentar falsamente la eficacia de cualquier intervención terapéutica, lo cual llevó a cuerpos colegiados a hacer repetidas exhortaciones a investigadores y editores para que publiquen los resultados de estudios en los que un medicamento no es diferente al placebo.

Correspondencia:

Alvar Loria
E-mail: alvarloria@gmail.com

Fecha de recepción: 04-08-2017
Fecha de aceptación: 13-02-2018
DOI://dx.doi.org/10.24875/GMM.18003627

Gac Med Mex. 2018;154:409-412
Disponible en PubMed
www.gacetamedicademexico.com

Primera afirmación

Un tamaño de muestra inadecuado tiene repercusiones importantes, no solo éticas y económicas sino también en interpretación y conclusiones de estudios clínicos.

En las estadísticas anuales de protocolos sometidos a los comités de investigación y ética del Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, el cálculo de TM siempre ha ocupado el primer lugar en fallas. Las principales son una información insuficiente para recalcularse TM (la más frecuente), el uso de ecuaciones obsoletas que no toman en cuenta el error beta y la falta de correspondencia entre la ecuación y el diseño del estudio (la menos frecuente).

No se salvan de problemas los investigadores de primer mundo ni los protocolos apoyados por la industria farmacéutica, como ejemplo: en 2009, solo una minoría, 188 de 446 (42%) de los protocolos sometidos a comités éticos del Reino Unido aportó una información completa que permitió recalcularse TM y no hubo gran diferencia entre protocolos apoyados por la industria y los no comerciales (44 versus 40%, respectivamente). Por otra parte, tener datos completos no resultó garantía ya que en 54 de los 188 completos (29%) se obtuvo un TM diferente al recalcularlo.²

Segunda afirmación

La falla más frecuente de los protocolos es que no proporcionan todos los datos para recalcularse TM.

El cálculo de TM no es perfecto, ya que requiere datos de estudios previos del mismo medicamento o de la misma intervención que se pretende estudiar. Está implícito que los resultados observados en investigaciones anteriores van a ser los mismos que el investigador va a observar en el suyo, lo cual es muy poco probable. Entra así un grado de incertidumbre en el cálculo de TM, pero es peor hacer estudios sin tener una idea, aunque sea aproximada, de cuántos casos se necesitan para considerar que una intervención es positiva, o bien, que no tiene ningún efecto.

Datos infaltables en cálculo de TM de protocolos

Son cuatro los datos a incluir en los protocolos de estudios clínicos controlados:

1. *Z alfa*: Es el valor Z que corresponde a un error alfa dado. El error alfa es rechazar falsamente H_0 [hipótesis nula]. Se usa el término *precisión* para referirse a la probabilidad de cometer este error.

2. *Z beta*: Es el valor Z que corresponde a un error beta dado. El error beta es aceptar falsamente H_0 . Se usa el término *poder* para referirse a la probabilidad de cometer el complemento de este error (1 – error beta), que puede ser visto como la probabilidad de no cometer este error, o bien, como la probabilidad de rechazar correctamente la H_0 .

3. *VAR*: Es la varianza de la variable de desenlace (*outcome*). La varianza es una medida de la variabilidad de una variable y se calcula diferente para datos numéricos ($VAR = \text{desviación estándar al cuadrado}$) y categóricos ($VAR = p \times q$). La muestra será mayor mientras mayor sea VAR.

4. *DIF*: Es la diferencia mínima en la variable de desenlace que los investigadores consideren clínicamente relevante. En los protocolos de estudio de dos grupos, control y tratado, se usa la diferencia de medias o de proporciones. El tamaño será mayor mientras menor sea DIF.

A la variable de desenlace se le dan nombres diversos, entre ellos, variable principal, variable de seguimiento, variable dependiente, pero prefiero usar *desenlace* como equivalente de *outcome*, lo cual está acorde con el hecho de que la variable de desenlace puede ser numérica (cuantitativa) o categórica (cualitativa).

Tercera afirmación

Se deben considerar obsoletas las ecuaciones que no tengan las dos zetas y no deben usarse ya que no consideran la Z del error beta.

De hecho, evaluar una falta de efecto se ha vuelto más importante que establecer un efecto positivo. Es decir, el racional del cálculo actual es opuesto a las primeras estrategias de cálculo de TM en las cuales era importante la probabilidad de que el tratamiento fuera positivo y ni siquiera se pensaba en qué tan negativo era un estudio negativo. Un ejemplo de este giro son las estrategias para evaluar estudios de dosis-respuesta que buscan un TM que permita alcanzar un poder de 90 o 95% y no toman en cuenta la precisión, dado que el interés se centra en establecer con certeza si hay o no diferencias entre dosis.

Cuarta afirmación

Actualmente, en las ecuaciones de cálculo de TM es tan importante el poder como la precisión.

Veamos un poco más sobre los cuatro componentes de las ecuaciones:

Z alfa y Z beta dependen de la voluntad del investigador dentro de ciertos límites, por ejemplo, es casi universalmente aceptado que la precisión esté entre 1 y 5 % y el poder entre 80 y 95 %. El rango de valores Z alfa y Z beta se presentan en las columnas 3 y 6 de la Tabla 1, usados por casi todos los investigadores clínicos, y de los cuales aconsejo no salirse si se desea que el protocolo sea aprobado.

Los valores Z más bajos de la tabla son los que darían el TM más bajo, pero se recomienda no usar valores Z más bajos, sobre todo del poder, lo cual ha llevado a diversos autores a recomendar el uso de poder mínimo de 90 o 95 %. Vea la NOTA encuadrada que sigue si quiere saber por qué hablar de valores Z es hablar de probabilidades.

Nota

Toda distribución de valores, sea gaussiana o no, se puede transformar a una distribución Z y al hacerlo, toda distribución Z tiene Media = cero y desviación estándar = uno.

Esto permite comparar las distribuciones de diferentes variables pues todas están en unidades DesvEst en lugar de las unidades en que se midieron. En consecuencia, los valores Z significan lo mismo para cualquier distribución Z. Así, Z alfa = 1.96 significa que 5 % del área quedará fuera a una distancia de 1.96 unidades de la media cero, y, por tanto, habrá 5 % de probabilidades de cometer un error alfa en cualquier distribución Z. Por eso, si se escoge precisión de 5 % se usa Z alfa = 1.96, pero si se escoge otra precisión, por ejemplo 1 %, tendrá que usarse Z alfa = 2.58 (ver Tabla 1).

En forma similar, Z beta = 0.84 deja fuera a 20% del área, o sea, las probabilidades de cometer un error beta es de 20 % (poder de 80 %), pero si se desea que el error beta sea menor (mayor poder), podría usarse Z beta = 1.28, que corresponde a 90 % de poder.

En un protocolo basta poner la precisión y el poder usados en la ecuación, por ejemplo, se calculó con precisión de 5 % y poder de 90 %.

VAR y DIF

Son los componentes que el investigador obtiene de estudios previos en que se aportan datos de respuesta a tratamiento de la misma intervención sobre la misma variable de desenlace. La solidez de estos componentes será mayor mientras mayor sea la información previa, y será mejor si hay información de la

Tabla 1. Rango de valores Z empleados con mayor frecuencia

Error alfa	Precisión	Z alfa	Error beta	Poder	Z beta
0.1 %	0.1 %	3.29	5 %	95 %	1.65
1 %	1 %	2.58	10 %	90 %	1.28
2 %	2 %	2.33	15 %	85 %	1.04
5 %	5 %	1.96	20 %	80 %	0.84

respuesta tanto en tratados como en controles (y no solo en tratados).

Los componentes Z alfa, Z beta y VAR siempre están en el numerador de las ecuaciones de TM en tanto que DIF está en el denominador. Por ello, para bajar TM solo queda disminuir VAR o aumentar DIF, ya que se recomienda no usar valores Z debajo de 1.96 y 0.84.

Lo ideal de un protocolo es que contenga una pregunta de investigación que se pueda contestar midiendo el efecto de una intervención sobre una sola variable de desenlace. Pero si se incluye más de una variable de desenlace, entonces debe calcularse TM para cada una y usar el TM mayor de los calculados.

Quinta afirmación

No se puede recalcular el TM si falta alguno de los cuatro componentes del cálculo. Si hay más de un desenlace, debe optarse por el TM mayor.

Los únicos protocolos sin componentes de cálculo son los estudios exploratorios, en los que no hay estudios previos. En tales casos se recomienda un estudio piloto con 10 a 20 sujetos que sirva para calcular TM. Esta estrategia es similar a la de los diseños adaptativos (*adaptive designs*), que permiten modificar un protocolo original, entre ellos TM, una vez que se tienen los primeros datos de un estudio.

TM en los ENI (estudios de no inferioridad)

Los avances terapéuticos han llevado a que sea una falta ética comparar nuevos medicamentos contra placebo, por lo que actualmente se recomienda comparar *versus* el tratamiento que se acepta como el mejor disponible hasta ese momento y que muchos denominan “control activo”, para recordarnos que va a actuar como si fuera el placebo del grupo control.

En los ENI se calcula el IC 95 % (intervalo de confianza) de la diferencia de medias o de proporciones de los dos grupos. El límite inferior del IC 95 % no debe rebasar el llamado margen de no inferioridad

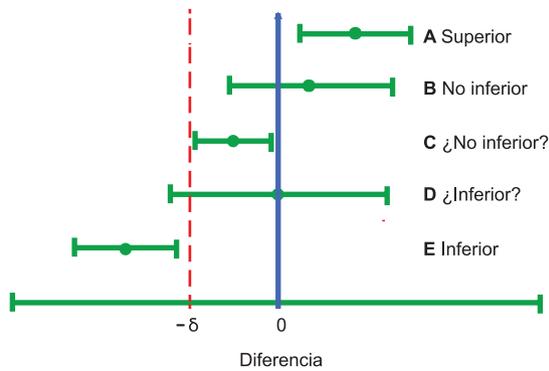


Figura 1. IC 95 % de cinco estudios (A-E). Diferencia de tratamiento nuevo menos control activo.

(simbolizado habitualmente por delta minúscula, δ). Si se grafica el IC 95 % a lo largo del eje X se facilita conocer los diferentes resultados que pueden darse en los ENI, por ejemplo, la Figura 1 muestra los IC 95 % de cinco estudios (A-E) y los clasifica de acuerdo a si el límite inferior de cada intervalo rebasa o no las verticales en cero y $-\delta$. De acuerdo con esto, los estudios A, B y E son concluyentes, en tanto que los estudios C y D necesitan incluir más sujetos de estudio para aclarar si son o no inferiores al control activo. El investigador escoge el margen (δ) basado en su juicio y en lo que se conoce del tema en la literatura médica. Hasta ahora existen pocas recomendaciones sobre maneras de establecer la magnitud de δ .³

Cálculo de TM en los ENI

El creciente empleo de los ENI amerita acotar que se pueden usar las mismas ecuaciones que en los

estudios de superioridad, solo que modificando el denominador de las ecuaciones. Basta restar el valor absoluto de δ a la diferencia de medias (o de proporciones) en el denominador de las ecuaciones:

$$\text{Denominador de no inferioridad} = (\text{DIF medias} - \delta)^2$$

$$\text{Denominador de superioridad} = (\text{DIF medias})^2$$

Puede verse que los ENI tendrán siempre un denominador menor que los de superioridad y, por lo tanto, el TM de un protocolo ENI será siempre mayor al de un protocolo de superioridad.

Conclusiones

El tamaño de efecto (*effect size*) tiene un papel importante en el cálculo de TM. Es una medida de la fuerza de un fenómeno y hay formas muy diversas de medirlo de modo que una correlación entre dos variables, un coeficiente de regresión, una diferencia de medias e, incluso, el riesgo de que algo suceda como qué proporción de sujetos sobrevivirá un ataque cardiaco, son ejemplos de cómo medimos un efecto.

La creciente sofisticación en la medición de tamaño de efecto conducirá a nuevas maneras de calcular TM.

Bibliografía

1. Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *J R Soc Med.* 2011;104:532-538.
2. Clark T, Berger U, Mansmann U. Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: review. *Brit Med J.* 2013;346:f1135.
3. Flight L, Julious SA. Practical guide to sample size calculations: non-inferiority and equivalence trials. *Pharmaceutical Stat.* 2016;15:80-88.