Parte 3 - **Claves para la interpretación de conceptos estadísticos en estudios de investigación**

Keys issues for interpretation of statistical concepts in clinical research studies

Bottaro FJ

Servicio de Clínica Médica - Hospital Británico
Coordinador Comité de Revisión Institucional - Hospital Británico

febottaro@hotmail.com

Fecha de recepción: 30/11/2013
Fecha de aprobación: 04/12/2013

COMO SE LEE UN ARTÍCULO CIENTÍFICO

HEMATOLOGÍA, Vol.17 N° 3: 299-305
Septiembre - Diciembre 2013

Palabras clave: riesgo relativo, número necesario a tratar, odss ratio, intervalo de confianza

Keywords: relative risk, odds ratio, number needed to treat, confidence interval

"Medicine is a science of uncertainty, and an art of probability" William Osler (1849-1919)

Ya sea que usted ame u odie la estadística, será necesario que cuente con cierta comprensión de algunos conceptos para poder evaluar críticamente la literatura médica. Para conseguir esto no es necesario saber cómo hacer un análisis estadístico, pero si será indispensable que conozca cómo interpretar la descripción de los resultados de un estudio y sus figuras. En este artículo de la serie abordaremos algunos conceptos básicos que son claves para la adecuada comprensión de los estudios de investigación.

1. Riesgo

Riesgo, en terminología estadística, puede ser definido como la probabilidad de ocurrencia de un evento futuro en una población determinada; y puede ser estimado dividiendo el número de personas que sufrieron un evento por el número total de personas inicialmente que no poseían el evento. En términos prácticos, cada vez que nos referiremos a riesgo estaremos hablando de un cociente o proporción. Por ejemplo, el riesgo de muerte por gripe durante el año 2013 se estimaría dividiendo el número de muertes secundarias a esta afección en el año por el número total de individuos que estaban en riesgo de padecer gripe al inicio del mismo.

En los estudios de investigación se puede comparar el riesgo entre dos poblaciones: una de ellas recibiendo un tratamiento estándar y la segunda un nuevo tratamiento o tratamiento experimental. El efecto de la nueva intervención puede ser expresado como el cambio relativo de sufrir un evento dado el nuevo tratamiento. Esto es denomina riesgo relativo (RR). La **figura 1** muestra la forma matemática de calcularlo. De acuerdo al ejemplo de la figura 1 podríamos decir que el RR de R-CHOP (tratamiento

experimental) sobre el esquema CHOP (tratamiento estándar) en pacientes ancianos con linfoma difuso células grandes B es 0,43. Es decir que los pacientes que recibieron esquema R-CHOP tuvieron menos de la mitad (43%) de riesgo de progresión de enfermedad que los pacientes que recibieron esquema CHOP; o también, que es lo mismo, el esquema R-CHOP disminuyó 57% el riesgo de tener progresión de enfermedad si se suma rituximab al esquema terapéutico convencional con CHOP (**Figura 2**).

Supongamos ahora un segundo estudio que evalúa el efecto de rituximab asociado a esquema CHOP en pacientes con linfoma difuso células grandes B pero con un riesgo de progresión más bajo que el ejemplo anterior (**Figura 2**). La reducción de riesgo relativo (RRR) con el rituximab en el Caso 2 permanece igual (57%), pero debido a que la tasa de eventos basal (en el grupo control) en esta población es más baja la reducción absoluta de riesgo (RAR) es solamente 5,7%.

Cuadro A	Tratamiento experimental	Tratamiento estándar	Total
Evento Si	A	C	
Evento NO	B	D	
Total	A + B	C + D	

Cuadro B	Rituximab + CHOP	CHOP	Total
Progresión de linfoma	19	43	62
NO progresión de linfoma	183	154	337
Total	202	197	399

Figura 1: Tabla 2x2 de un estudio que compara esquema CHOP versus R-CHOP en pacientes ancianos con linfoma difuso células grandes B.1 En el cuadro A se observa el esquema general de una tabla 2x2 típica. En el cuadro B se muestra la tabla 2x2 para el punto final primario "Progresión de enfermedad de base" del estudio de referencia.
***Riesgo de progresión con CHOP:** c/c+d= 43/197= 0,22= 22 %*
***Riesgo de progresión con R-CHOP**: a/a+b= 19/202= 0,094= 9,4 %*
***Riesgo Relativo**= (a/(a+b))/(c/(c+d))= 0,094/0,22= 0,43*
***Reducción de Riesgo Relativo**= 1-[(a/(a+b))/(c/(c+d)]= 1 – 0,43= 0,57= 57%*
***Diferencia de riesgo o Reducción Absoluta de Riesgo**= (a/(a+b))-(c/(c+d))= – 0,22= -0,124*
***Número necesario a tratar:** 1/ [(a/(a+b))/(c/(c+d))]= 1/0,124= 8*
1. Coiffier B, Lepage E, Briere J, y col. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. N Engl J Med. 2002 Jan 24;346(4):235-42

A pesar que la RRR de una intervención puede ser similar en diferentes grupos de riesgo, la ganancia absoluta, representada por la RAR, puede modificarse sustancialmente alterando el impacto de la intervención. Cuando las tasas de eventos son bajas la RAR disminuye, mientras que la RRR o eficacia del tratamiento puede permanecer constante. Esta es la razón por lo que algunas intervenciones están justificadas en determinados grupos de riesgo de una enfermedad, y no estarlo en otros grupos, a pesar que la eficacia de la intervención no se ha modificado. Es notable remarcar como el número necesario a tratar (NNT) aumenta más del doble (de 8 en el estudio original a 18 en el hipotético Caso 2) en nuestro ejemplo de la figura 2 sin modificación del RR.

Por esta razón, podemos decir que las medidas relativas (RR y odds ratio) miden la "**fuerza** de la asociación" entre dos variables, mientras que las medidas absolutas estiman su **impacto**. Este fenómeno es bien conocido por la industria farmacéutica, utilizando esta situación a menudo para su beneficio. Por ejemplo, una droga es testeada en ensayos clínicos que incluyen poblaciones de alto riesgo de ocurrencia de eventos (pacientes muy enfermos) hallándose reducciones absolutas de riesgo llamativas, pero subsiguientemente es comercializada para su uso en poblaciones menos afectadas y en los que la RAR es mucho menor. Cuanto menor es la tasa de eventos en el grupo control de un estudio, mayor será la diferencia entre RRR y RAR.

La RRR siempre es más atractiva que la RAR, siendo también el principal parámetro de difusión comercial. Sin embargo, el impacto clínico y utilidad real de una intervención es más apreciable cuando consideramos las medidas absolutas.

Cuando queremos calcular una medida relativa de

la asociación entre dos variables cuyo punto final es calculado sobre una línea temporal (por ejemplo en los análisis de sobrevida, donde no solo nos importará si el punto final ha ocurrido o no; también será importante el tiempo transcurrido hasta la ocurrencia del evento) en lugar de calcular el RR deberíamos calcular el *"hazard ratio"* (HR). La interpretación del HR es similar al RR, es decir que un HR por arriba de 1 significa que existe un riesgo incrementado de eventos con la intervención o exposición, mientras que un valor por debajo del 1 implica un efecto protector. Un valor de 1 se interpreta como ausencia de efecto entre las intervenciones o ausencia de efecto de una exposición.

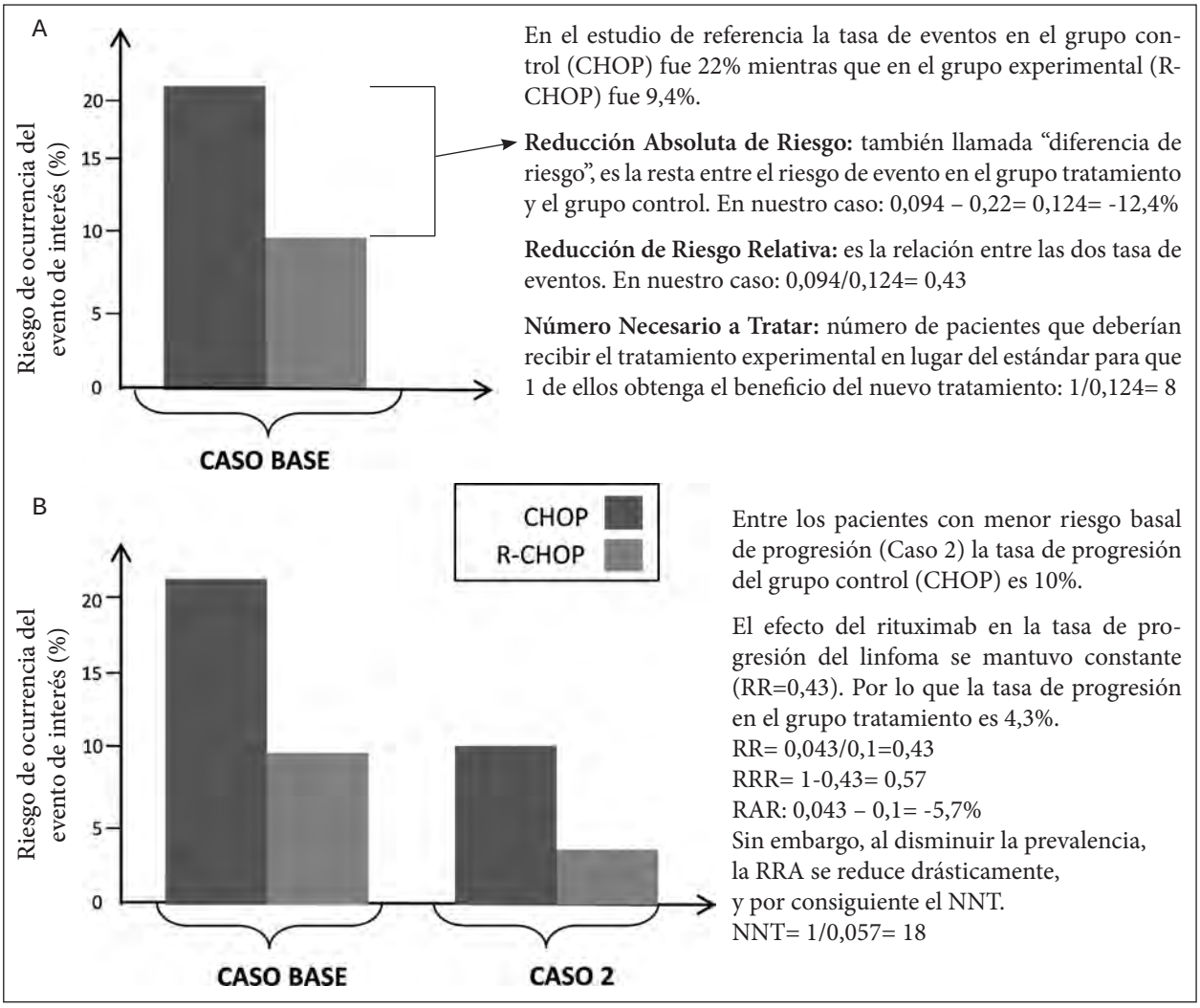

Figura 2: *Resultados del estudio de la figura 1. Las barras representan la tasa de progresión durante el tratamiento en los dos grupos del estudio.* ***A****: Resultados del estudio de referencia figura 1.* ***B****: Resultados de un estudio hipotético que comparte igual intervención, tratamiento control e igual punto final, pero la población que participa posee un riesgo de progresión de enfermedad más bajo que el caso base (****figura 1A****).*
RRR: reducción riesgo relativo. RAR: reducción absoluta de riesgo. RR: riesgo relativo. NNT: Número necesario a tratar

El *odds ratio* (OR) es también una medida de la fuerza de una asociación entre dos variables. Ha sido traducido de múltiples formas al castellano (razón de oportunidades, razón de probabilidades, o razón de productos cruzados son solo algunas) sin que ninguna de estas reemplazará el término inglés. Una buena opción que sirve para evitar confusiones y se ha hecho usual es incorporar directamente el término inglés y decir siempre *"odds ratio"*. Si bien conceptualmente su interpretación es similar al RR, su fórmula matemática (ver **Figura 4**) y algunas características particulares lo diferencian de este. El concepto de

"odds" (chance) se maneja habitualmente en el mundo anglosajón, en especial en las casas de apuestas y corresponde al cociente entre la chance de que un evento ocurra y la chance de que no ocurra.
Debido a que los denominadores entre RR y OR son diferentes, el OR solo nos puede dar una estimación aproximada de riesgo. Sin embargo, existen situaciones en donde no es adecuado utilizar el RR y deberemos usar solamente OR, como por ejemplo los estudios de casos y controles.

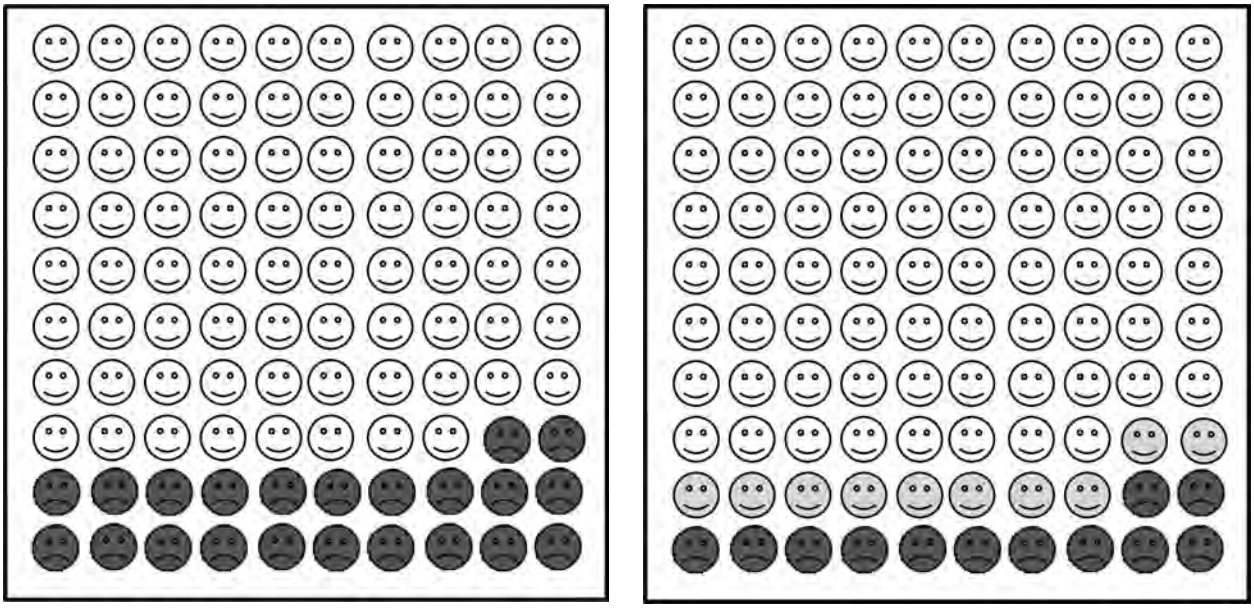

Figura 3: *representación conceptual de los resultados del estudio de referencia de* ***Figura 1****. En blanco pacientes que no presentaron progresión de enfermedad, en gris oscuro: pacientes que presentaron progresión de enfermedad. En gris claro: pacientes que de haber recibido el tratamiento estándar hubieran presentado progresión de su enfermedad. Los pacientes en gris claro representan la diferencia de riesgo o reducción de riesgo absoluta.*

Cuadro A	Exposición Si	Exposición No	Total
Evento Si	A	C	
Evento NO	B	D	
Total			

Cuadro B	Rituximab + CHOP	CHOP	Total
Progresión de linfoma	19	43	
NO progresión de linfoma	183	154	
Total			

Figura 4: *Tabla 2x2 que respeta los valores observados en el estudio de referencia 1 ilustrativa para cálculo de odds ratio*

Odds ratio (OR)*= (a/b)/(c/d)=(19/183)/(43/154)= 0,1/0,28= 0,36*

Riesgo Relativo*= (a/(a+b))/(c/(c+d))= 0,094/0,22= 0,43*

El OR es calculado como un cociente o relación entre dos "chances" u "odds". Es el cociente entre la chance de que un evento ocurra y la chance de que no ocurra. Para su cálculo se divide el número de personas expuestos a un factor (ejemplo tratamiento experimental) que sufrieron el evento sobre el número de personas que no lo sufrieron; dividido el número de personas que sufrieron el evento sin exposición a este factor sobre el número de personas que no lo sufrieron.

2. Número necesario a tratar (NNT)

El NNT es el número de pacientes que un médico debería tratar con un tratamiento en particular para prevenir que 1 paciente desarrolle un evento no deseado en un período de tiempo. Si por ejemplo el NNT para un tratamiento es 15, significa que deberíamos administrar a 15 pacientes el tratamiento para evitar que uno de ellos desarrolle un evento no deseado, es decir que cada paciente que reciba el tratamiento tendrá 1 en 15 chances de ser el beneficiario.

Cuanto mayor sea la RAR, menor será el NNT y por lo tanto deberemos tratar a muy pocos pacientes para observar un beneficiario. Si la RAR es muy pequeña, entonces el NNT indefectiblemente aumentará. El NNT es la inversa del RAR (ver Figura 2).

Un concepto similar al NNT es el número necesario para dañar (NNH). Y su interpretación es similar, es el número de pacientes que un médico debería tratar con un tratamiento en particular para que 1 paciente desarrolle un evento adverso secundario al tratamiento en un período de tiempo. El NNT y NNH son dos medidas del impacto de una asociación que pueden ser utilizados en la práctica diaria para valorar el beneficio-clínico neto. Los médicos habitualmente poseen dificultades para interpretar este concepto. Repasaremos de la literatura[1] una buena manera de reflexionar sobre este concepto: si una enfermedad posee una mortalidad del 100% sin tratamiento, y un tratamiento reduce la muerte a 50%, ¿cuántas personas deberían ser tratadas para prevenir una muerte? De los datos que brindamos podríamos decir que tratar a 100 personas con la enfermedad en cuestión resultaría en 50 sobrevivientes. Esto es equivalente a decir que 1 de cada 2 pacientes tratados "salvan" su vida. El mismo valor se obtiene al dividir 1 sobre 0,5; que sería la RAR de nuestro ejemplo.

3. Medidas de precisión

Al comparar la eficacia de dos intervenciones podemos utilizar el test de hipótesis (pruebas de significación estadística y su resultante valor de p) o técnicas de estimación que nos proporcionarán intervalos de confianza.[2]

Las pruebas de significación estadística calculan la probabilidad que los resultados observados en un estudio controlado (EC) puedan ser debidos al azar, en el supuesto de que ambas intervenciones fueran iguales en su eficacia (hipótesis nula cierta). Esta probabilidad es lo que habitualmente se denomina *"valor de p"*. Por consenso se acepta que un valor de p=0,05 como punto de corte por debajo del cual se considera que se puede rechazar la hipótesis de igualdad de ambas intervenciones y concluir que el resultado es estadísticamente significativo, es decir que interpretamos que las intervenciones son diferentes. Un valor de p<0,05 solo nos permitirá rechazar la hipótesis de igualdad (o nula) de dos intervenciones, pero no nos dará información alguna sobre la magnitud de la diferencia o el sentido de esta.

El test a través del cual llegamos a obtener el valor de p (test de significancia) variará de acuerdo a la naturaleza de las variables que se usaron para medir el efecto de las intervenciones (variable cuantitativa o cualitativa), su distribución (si estas son cuantitativas), la independencia de las poblaciones sobre las que fueron evaluadas las intervenciones (grupos paralelos, apareados, etc) (ver **Tabla 1**).

Los intervalos de confianza (IC), otro de los métodos que podemos utilizar para comparar la eficacia de dos intervenciones pueden resultarnos más útiles ya que nos dan una idea del tamaño o relevancia del efecto observado. Nos permiten conocer entre que valores límite se encuentre la verdadera diferencia.

Usualmente se calculan IC del 95%, lo que significa que un 95% de las veces el valor verdadero de la diferencia de eficacia entre dos intervenciones se encontrará en el rango de los valores delimitados por el IC. Por ejemplo, si encontraramos que la diferencia observada entre dos intervenciones es de 25% (p<0,05; IC 95% 19-31%), esto significa que podemos tener 95% de confianza de que la diferencia real entre las dos intervenciones se encuentra entre 19 y 31%.

Cuando la medida de efecto que analizamos es la diferencia de efecto entre dos intervenciones, deberemos tener en cuenta que si el IC incluye el valor 0 se concluye que el resultado no es estadísticamente significativo; es decir que no existen diferencias entre el efecto observado de los dos tratamientos. Si en lugar de una diferencia absoluta nos interesa una medida relativa como el RR, cuando el IC incluya el valor 1 concluiremos que ambos tratamientos son iguales y que por lo tanto no hay significación estadística.

Los IC nos brindan, además, información sobre cuán grande o pequeño puede ser el verdadero efecto de una intervención. Si el IC es estrecho podremos estar tranquilos que valores fuera de este rango han sido

descartados por el estudio, esto suele ocurrir cuando el tamaño de muestra de un estudio es muy grande y la estimación del verdadero efecto es muy precisa. Esto significa que el estudio posee "poder" suficiente para hallar una diferencia entre dos intervenciones. Si el estudio es muy pequeño (poco tamaño muestral) el IC será muy amplio, implicando un rango diverso de tamaños de efecto y por lo tanto brindando información imprecisa.

4. Potenciales errores en la interpretación

Tabla 1: Resumen de los test de significación estadística.

Diferencia que se quiere testear	Datos cuantitativos de distribución normal o paramétrica	Datos continuos de distribución no normal o no paramétricos/ datos númericos discretos/datos categóricos ordinales	Datos categóricos binarios
Una sola muestra	Test-t para un muestra		Test Chi^2 (o test exacto de Fisher)
Dos muestras independientes	Test-t	Test de la U o Mann-Withney	Test Chi^2 (o test exacto de Fisher)
Dos muestras apareadas	Test-t para muestras apareadas	Test de Wilcoxon para muestras apareadas	Test McNemar
Tres o más muestras independientes	ANOVA	Test de Kruskall-Wallis	Chi^2
Tres o más muestras apareadas	ANOVA	Test de Friedman	Test McNemar

de test estadísticos

Error tipo I o error alfa: este tipo de errores ocurre cuando en un EC rechazamos la hipótesis nula (igualdad entre las intervenciones) cuando en realidad esta es verdadera. En otras palabras, asumimos como diferentes dos tratamientos o intervenciones cuando en realidad su efecto es similar. Los test de significación contribuyen a resguardarnos de este error intentando determinar la variabilidad del tamaño del efecto con un margen de error pre-especificado de 5% (valor p=0,05). Por ejemplo, si la hipótesis nula es verdadera:

- Un valor p <0,05 significa que la probabilidad de obtener ese resultado por azar es menor a 5%
- Un valor p <0,025 significa que la probabilidad de obtener ese resultado por azar es menor a 2,5%
- Un valor p <0,01 significa que la probabilidad de obtener ese resultado por azar es menor a 1%

La determinación de 0,05 como punto de corte para el valor de p es completamente arbitraria y representa el mínimo grado de evidencia que se necesita en orden de descartar la hipótesis nula.
Esto significa que cuando el valor de p es igual a 0,05 existe una probabilidad que 1 de cada 20 resultados "significativos" podría ser falso y las diferencias encontradas serían fruto del azar. Esto es particularmente cuando se realizan múltiples comparaciones sobre la ocurrencia de un mismo evento. En estos casos el valor de p debería ser "corregido" y se debiera considerar disminuir el valor de corte del valor p.
Error tipo II o error Beta: otro posible error que podemos cometer es cuando consideramos que debido a un resultado “no significativo” no hay efecto, cuando en realidad este si existe. Este tipo de error es frecuente y muy perjudicial en muchas ocasiones. Un IC no significativo nos dice simplemente que la diferencia observada en nuestro estudio es consistente con que no hay verdadera diferencia entre los grupos. Pero debemos ser cuidadosos, porque solo porque no hemos hallado diferencias entre las intervenciones en nuestro estudio no implica necesariamente que estas no existan. La probabilidad de cometer un error de esta índole suele denominarse como β, y por lo tanto el poder de un estudio se calculará como 1-β. El poder de un estudio implica la capacidad de un estudio para hallar una diferencia cuando esta realmente existe. El poder de un estudio está estrechamente relacionado con el tamaño

muestral del estudio. Habitualmente los investigadores suelen establecer la magnitud mínima de la diferencia o asociación que se considera de relevancia clínica, así como el poder estadístico que se desea para el estudio y, de acuerdo a estos parámetros, calcular el tamaño de la muestra necesaria.

5. Relevancia clínica y significación estadística

La significación estadística muchas veces se interpreta de manera incorrecta al asociarla con un resultado importante. Los test de significación estadística solo podrán ayudarnos a determinar si los datos que proporciona un estudio pueden ser debidos al azar o no. Un estudio de muy grandes dimensiones (elevado tamaño muestral) puede hallar como estadísticamente significativas diferencias pequeñas que poseerán escasa relevancia clínica. Es el tamaño del efecto, no el tamaño de la significación (cuan chico es el valor de p) lo que es relevante para la práctica clínica.

Para evaluar la relevancia de los resultados de un estudio se deberán reportar la RRR, RAR, NNT.

6. Análisis post-hoc

Se denomina así a los análisis estadísticos que se realizan al concluir un estudio de investigación que no estaban pre-especificados antes del inicio del estudio. Muchas veces se seleccionan aquellos pacientes en los que la intervención ha resultado más eficaz y se los agrupo intentando buscar alguna característica común. Un interesante ejemplo en la literatura es el estudio CHARISMA, un estudio clínico aleatorizado controlado que randomizó casi 15000 pacientes con enfermedad cardiovascular establecida o múltiples factores de riesgo a recibir aspirina más clopidogrel versus aspirina más placebo por una media de 28 meses. El punto final fue la combinación de infarto de miocardio, accidente cerebrovascular y mortalidad de causa cardiovascular.[3] Llamativamente los autores del estudio concluyeron que se evidenció una "sugerencia" de beneficio del clopidogrel en el subgrupo de pacientes con aterotrombosis sintomática, un subgrupo especial de pacientes creado al final del estudio. Esto origino severas críticas, incluyendo la propia editorial del estudio publicada en el New England Journal of Medicine.[4, 5]

Los resultados de los análisis post-hoc no son válidos para demostrar o refutar la hipótesis del estudio, solo sirven para generar otras hipótesis que deberán ser demostradas adecuadamente en otros estudios diseñados para tal fin. También pueden ser útiles en la generación de alertas de seguridad en base a datos de farmacovigilancia.

Declaración de conflictos de intereses

El autor declara no poseer conflictos de interés.

Bibliografía

1. Barratt A, Wyer PC, Hatala R; Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. CMAJ. 2004 Aug 17;171(4):353-8.

2. Altman DG, Machin D, Bryan TN, y col. Staistics with confidence. Bristol: JW Arrowsmith; 2002.

3. Bhatt DL, Fox KA, Hacke W, y col; CHARISMA Investigators. Clopidogrel and aspirin versus aspirin alone for the prevention of atherothrombotic events. N Engl J Med. 2006 Apr 20;354(16):1706-17.

4. Pfeffer MA, Jarcho JA. The charisma of subgroups and the subgroups of CHARISMA. N Engl J Med. 2006;354(16):1744-6.

5. Gebel JM Jr. The CAPRIE-like subgroups of CHARISMA: a CAPRIEciously biased analysis of an unCHARISMAtic truth. J Am Coll Cardiol. 2007; 50(17):170