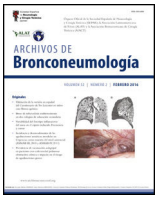
Arch Bronconeumol. 2020;56(9):547–548

Sociedad Española de Neumología y Cirugía Torácica SEPAR

ARCHIVOS DE Bronconeumología

www.archbronconeumol.org



Editorial

La ruleta rusa de los análisis estadísticos de estudios clínicos. Una reflexión para lectores y autores

The Russian Roulette of Statistical Analysis in Clinical Trials. Some Considerations for Readers and Authors

Los ensayos clínicos –y también los estudios observacionales aplicados a clínica– están diseñados para responder una o varias preguntas de investigación. La pregunta de investigación lleva implícita una o varias exposiciones (tratamiento, hábito de salud) y uno o varios efectos en salud, que se formulan en forma de hipótesis de investigación. Para contrastar estas hipótesis se diseña un estudio, que al ejecutarlo se obtienen unos datos. En la fase de análisis, esos datos se contrastan con las hipótesis inicialmente planteadas, aplicando las pruebas estadísticas adecuadas según la naturaleza de las variables independientes y dependientes.

Si en el análisis estadístico se observan diferencias en la variable respuesta (según el nivel de la exposición; p. ej. tratamientos), estas pueden deberse a: 1) el efecto real del tratamiento, 2) un sesgo, o/y 3) a un error aleatorizado (azar)[1]. La estadística nos permite cuantificar qué probabilidades hay de que las diferencias encontradas entre las exposiciones se deban a un error aleatorizado en el proceso de seleccionar la muestra de estudio. Debemos pensar que los sujetos que entran a formar parte del estudio son en realidad una muestra de la población a estudio, y que de esta población se pueden extraer infinitas muestras, que van a ser diferentes entre ellas, y potencialmente distintas a la población. Por tanto, la muestra de nuestro estudio puede diferir aleatoriamente de la población.

Mediante los test de hipótesis, la estadística calcula la probabilidad de que, no existiendo esas diferencias entre los tratamientos en la población, las hayamos encontrado en nuestros datos por azar. Si esta probabilidad es menor que una fijada a priori (riesgo alfa = 0,05), podemos concluir con una seguridad del 95% que esta diferencia también existe en la población. Pero con ello estamos admitiendo que nos podemos equivocar una de cada 20 veces concluyendo que existen diferencias en la población, cuando realmente no las hay (falso positivo).

El problema

El problema surge cuando realizamos múltiples comparaciones no planificadas. Cuantas más comparaciones hagamos, mayor será la probabilidad de encontrar que alguna de ellas sea un falso positivo[2]. Si (*NC*) es el número de comparaciones, la probabilidad de encontrar al menos una asociación estadísticamente significativa por azar es: $1-(0{,}95)^{NC}$. Así, si se realizan 25 comparaciones, la probabilidad de encontrar un falso positivo pasa del 5% al 72%. Y esto puede ser especialmente problemático cuando se hace una publicación selectiva de los resultados estadísticamente significativos,[3] ya que la probabilidad de que se deban al azar es cercana a uno.

Principales modos de realizar comparaciones múltiples en estudios clínicos

Incluir nuevas variables respuesta

La pregunta de investigación de un ensayo, generalmente, lleva implícita una única variable respuesta principal, pero es muy común que los investigadores analicen otras que no estaban planificadas, sobre todo cuando los resultados con la variable respuesta inicial no es el esperado al plantear la hipótesis. Se buscan otras variables, o se construyen unas nuevas a partir de las existentes para «ver si algo sale estadísticamente significativo». Ello aumenta el riesgo de encontrar falsos positivos.

Para evitarlo, los autores deben evitar incluir variables resultado no planificadas y, si lo hacen, dejar claro que es un análisis *post hoc* (a posteriori), además de mostrar todas las comparaciones realizadas[2]. Esto evita realizar una publicación selectiva de los resultados estadísticamente significativos. Para evitar una interpretación sesgada por la publicación selectiva, los lectores, revisores y editores deberían comprobar que las variables resultado estaban contempladas *a priori* en la metodología y en el registro del ensayo. Existen múltiples registros de ensayos[4–6], pero en ocasiones (dependiendo del país y el año de realización) puede ocurrir que el grado de detalle sobre la construcción de las variables resultado sea insuficiente.

También la categorización de las variables respuesta está sujeta a selección oportunista *post hoc* en función de los resultados más llamativos. Por ello, sería interesante establecer en el protocolo los puntos de corte o usar unos de referencia internacionales ya admitidos[7].

Análisis por subgrupos

A diferencia del análisis por grupos (definido a priori por la pregunta de investigación), en el análisis por subgrupos, estos suelen

https://doi.org/10.1016/j.arbres.2019.10.024

definirse a posteriori (una vez finalizada la recogida de datos). En el análisis por subgrupos se evalúan los efectos del tratamiento en subpoblaciones de pacientes según características basales clínicas (p. ej. diabetes, gravedad) o personales (sexo, edad, genética). El objetivo es evaluar si los efectos de los tratamientos varían dependiendo de ciertas características de los pacientes.

Existe una controversia sobre si se deben realizar o no[8]: no hacerlos puede privar de nuevos conocimientos científicos, y realizarlos aumenta la posibilidad de obtener falsos positivos por la utilización de comparaciones múltiples.

Las recomendaciones para tener en cuenta en los análisis por subgrupos son[9]:

(1) En el caso de que el análisis por subgrupos estuviera establecido en el protocolo, y su número sea reducido.
(2) Utilizar cálculos de interacciones para valorar la heterogeneidad entre grupos. El nivel de significación estadística de las interacciones permite valorar si el azar puede explicar las diferencias entre los subgrupos.
(3) Los resultados son consistentes con otros estudios y son biológicamente plausibles.

Análisis intermedios

Son análisis de datos realizados antes de finalizar el estudio, y tienen por objetivo obtener resultados preliminares que aconsejen interrumpir o no el estudio. Los principales motivos para detener un estudio son que[9]: 1) se determine que un medicamento es dañino, 2) sea altamente beneficioso, 3) la probabilidad de demostrar eficacia sea insignificante. Esta decisión la debería tomar un comité de seguimiento de datos independiente y no los propios investigadores. En todo caso, es necesario resaltar que aun en este caso, y a pesar de que se corrija el nivel de significación por técnicas estadísticas como O'Brien-Fleming, Peto o Pocock,[9] las estimaciones del efecto del tratamiento estarán siempre sesgadas.

Otros aspectos metodológicos que pueden aumentar el error tipo I de un ensayo son cómo tratar los valores ausentes (hay distintos métodos estadísticos para hacerlo), o cómo analizar las pérdidas durante el seguimiento (intención de tratar o por protocolo).

Conclusión

Existen múltiples maneras de hacer comparaciones no planificadas: con distintas variables respuesta, con distintas formas de agruparlas o categorizarlas, con análisis intermedios, realizando análisis por subgrupos, con distintas maneras de abordar los datos ausentes o las pérdidas durante el seguimiento.

Todas estas posibilidades no son aisladas, sino que se pueden dar combinaciones de ellas, lo que hace que las posibilidades de comparaciones no planificadas se multipliquen. Así, en el caso conservador de solo probar dos variables respuesta (*VR*), con dos puntos de corte (*PC*), con un análisis intermedio y uno final (*AI*), y dos análisis por subgrupos (*AS*), probando dos maneras de tratar los datos ausentes (*DA*), el número final de combinaciones es de:

$$NC = 2VRx2PCx2AIx2ASx2DA = 32$$

De esta manera, la probabilidad de encontrar un falso positivo asciende del 5% al 81% $\left(1 - 0.95^{NC} = 0.81\right)$. Y, si el autor solo muestra en el artículo los «estadísticamente significativos», las probabilidades de que ese hallazgo sea un falso positivo es altísima. Y esto puede ser especialmente grave para la salud pública y en clínica si las decisiones se basan en estos falsos positivos.

Por ello, es especialmente importante que todos los análisis estadísticos estén planificados a priori, y que en los registros de los ensayos se detallen todos los análisis que se van a realizar. Así editores, revisores y lectores pueden cotejar lo publicado con lo planificado, lo que les permitirá valorar hasta qué punto el resultado obtenido se puede deber a un falso positivo.

Bibliografía

1. Figueiras A. Causalidad en epidemiología. In: Hernández-Aguado I, Lumbreras B, editors. Manual de Epidemiología y Salud Pública para Grados en Ciencias de la Salud, 3.ª ed., España; 2018.
2. Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. Lancet. 2005;365:1591–5.
3. Nissen SB, Magidson T, Gross K, Bergstrom CT. Publication bias and the canonization of false facts. Elife. 2016;5:e21451, http://dx.doi.org/10.7554/eLife.21451.
4. Registro Español de Estudios Clínicos. Agencia Española de Medicamentos y Productos Sanitarios [consultado 9 Ene 2020]. Disponible en: https://reec.aemps.es/reec/public/web.html.
5. ClinicalTrials.gov. U.S. National Library of Medicine [consultado 9 Ene 2020]. Disponible en: https://clinicaltrials.gov/.
6. ISRCTN registry. BioMed Central, part of Springer Nature [consultado 9 Ene 2020]. Disponible en: https://www.isrctn.com/.
7. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemp Clin Trials Commun. 2018;11:156–64, http://dx.doi.org/10.1016/j.conctc.2018.08.001.
8. Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. J Clin Epidemiol. 1999;52:19–26.
9. Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet. 2005;365:1657–61.

María Piñeiro-Lamas[a,b], Margarita Taracido[a,b,c] y Adolfo Figueiras[a,b,c,*]

[a] *Consorcio de Investigación Biomédica en Epidemiología y Salud Pública (CIBER en Epidemiología y Salud Pública), Santiago de Compostela, España*

[b] *Fundación Instituto de Investigación Sanitaria de Santiago de Compostela (FIDIS), Santiago de Compostela, España*

[c] *Departamento de Medicina Preventiva y Salud Pública, Universidad de Santiago de Compostela, Santiago de Compostela, España*

* Autor para correspondencia.
Correo electrónico: adolfo.figueiras@usc.es (A. Figueiras).