ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Febrero 2020

FORMACIÓN MÉDICA

El detector de tramposos. Usos incorrectos de la estadística.

Molina Arias M.

Hospital Infantil Universitario La Paz.

Resumen

Una alta proporción de los trabajos publicados en revistas médicas son defectuosos desde el punto de vista metodológico. Estos errores pueden ser debidos a la falta de formación de autores y revisores, eminentemente clínicos. Sin embargo, en algunas ocasiones son errores deliberados con el objetivo de favorecer la obtención de determinadas conclusiones. Se revisan los errores más frecuentes que pueden observarse con el empleo de las pruebas estadísticas.

Introducción

Una alta proporción de los trabajos publicados en revistas médicas son defectuosos desde el punto de vista metodológico. Estos errores pueden ser debidos a la falta de formación de autores y revisores, eminentemente clínicos. Sin embargo, en algunas ocasiones son errores deliberados con el objetivo de favorecer la obtención de determinadas conclusiones. Se revisan los errores más frecuentes que pueden observarse con el empleo de las pruebas estadísticas.

Cuando pensamos en inventos e inventores, a la mayoría de nosotros nos viene a la cabeza el nombre de Thomas Alva Edison, conocido entre sus amigos como el mago de Menlo Park. Este señor creó más de mil inventos, de algunos de los cuales puede decirse que cambiaron el mundo. Entre ellos podemos nombrar la bombilla incandescente, el fonógrafo, el kinetoscopio, el polígrafo, el telégrafo cuádruplex, etc., etc., etc. Pero quizás su gran mérito no sea el de haber inventado todas estas cosas, sino el de aplicar métodos de producción en cadena y de trabajo en equipo al proceso de investigación, favoreciendo la difusión de sus inventos y la creación del primer laboratorio de investigación industrial.

Pero a pesar de toda su genialidad y excelencia, a Edison se le pasó inventar algo que habría tenido tanta utilidad como la bombilla: un detector de tramposos. La explicación de esta falta es doble: vivió entre los siglos XIX y XX y no se dedicaba a leer artículos sobre medicina. Si hubiese vivido en nuestro tiempo y hubiese tenido que leer literatura médica, no me cabe duda de que el mago de Menlo Park se habría dado cuenta de la utilidad de este invento y se habría puesto las pilas (que,

por cierto, no las inventó él, sino Alessandro Volta).

Y no es que yo esté hoy especialmente negativo, el problema es que, como ya dijo Altman hace más de 15 años, el material remitido a las revistas médicas es malo desde el punto de vista metodológico en un altísimo porcentaje de los casos. Es triste, pero el sitio más adecuado para guardar muchos de los trabajos que se publican es el cubo de la basura.

En la mayor parte de los casos la causa probablemente sea la ignorancia de los que escribimos. "Somos clínicos", nos decimos, así que dejamos de lado los aspectos metodológicos, de los cuales tenemos una formación, en general, bastante deficiente. Para arreglarlo, las revistas mandan revisar nuestros trabajos a otros colegas, que andan más o menos como nosotros. "Somos clínicos", se dicen, así que se comen todos nuestros errores.

Aunque esto es, de por sí, grave, puede tener remedio: estudiar. Pero es un hecho todavía más grave que, en ocasiones, estos errores pueden ser intencionados con el objetivo de inducir al lector a llegar a una determinada conclusión tras la lectura del trabajo. El remedio para este problema es hacer una lectura crítica del trabajo, prestando atención a la validez interna del estudio. En este sentido, quizás el aspecto más difícil de valorar para el clínico sin formación metodológica sea el relacionado con la estadística empleada para analizar los resultados del trabajo. Es aquí, sin ninguna duda, donde mejor se pueden aprovechar de nuestra ignorancia utilizando métodos que proporcionen resultados más vistosos, en lugar de los métodos adecuados.

Como sé que no vais a estar dispuestos a hacer un máster sobre bioestadística, en espera de que alguien invente el detector de tramposos, vamos a dar una serie de pistas para que el personal no experto pueda sospechar la existencia de estas trampas.

La primera puede parecer una obviedad, pero no lo es: **¿se ha utilizado algún método estadístico?** Aunque es excepcionalmente raro, puede haber autores que no consideren utilizar ninguno. Recuerdo un congreso al que pude asistir en el que se exponían los valores de una variable a lo largo del estudio que, primero, subían y, después, bajaban, lo que permitía concluir que el resultado no era "muy allá". Como es lógico y evidente, toda comparación debe hacerse con el adecuado contraste de hipótesis e indicarse su nivel de significación y la prueba estadística utilizada. En caso contrario, las conclusiones carecerán de validez alguna.

Un aspecto clave de cualquier estudio, especialmente en los de intervención, es el cálculo previo del **tamaño muestral necesario**. El investigador debe definir el efecto clínicamente importante que quiere ser capaz de detectar con su estudio y calcular a continuación qué tamaño muestral le proporcionará al estudio la potencia suficiente para demostrarlo. La muestra de un estudio no es grande o pequeña, sino suficiente o insuficiente. Si la muestra no es suficiente, puede no detectarse un efecto existente por falta de potencia (error de tipo 2). Por otro lado, una muestra mayor de lo necesario puede mostrar como estadísticamente significativo un efecto que no sea relevante desde el punto de vista clínico. Aquí hay dos trampas muy habituales. Primero, el del estudio que no alcanza significación y sus autores afirman que es por falta de potencia (por tamaño muestral insuficiente), pero no hacen ningún esfuerzo por calcular la potencia, que siempre puede hacerse a posteriori. En ese caso, podemos hacerlo nosotros

usando programas de estadística o cualquiera de las calculadoras disponibles en internet, como la GRANMO. Segundo, se aumenta el tamaño muestral hasta que la diferencia observada sea significativa, encontrando la ansiada $p < 0,05$. Este caso es más sencillo: solo tenemos que valorar si el efecto encontrado es relevante desde el punto de vista clínico. Os aconsejo practicar y comparar los tamaños muestrales necesarios de los estudios con los que definen los autores. A lo mejor os lleváis alguna sorpresa.

Una vez seleccionados los participantes, un aspecto fundamental es el de la **homogeneidad de los grupos basales**. Esto es especialmente importante en el caso de los ensayos clínicos: si queremos estar seguros de que la diferencia de efecto observada entre los dos grupos se debe a la intervención, los dos grupos deben ser iguales en todo, menos en la intervención.

Para esto nos fijaremos en la clásica tabla I de la publicación del ensayo. Aquí tenemos que decir que, si hemos repartido los participantes al azar entre los dos grupos, cualquier diferencia entre ellos se deberá, sí o sí, al azar. No os dejéis engañar por las p, recordad que el tamaño muestral está calculado para la magnitud clínicamente importante de la variable principal, no para las características basales de los dos grupos. Si veis alguna diferencia y os parece clínicamente relevante, habrá que comprobar que los autores han tenido en cuenta su influencia sobre los resultados del estudio y han hecho el ajuste pertinente durante la fase de análisis.

El siguiente punto es el de la **aleatorización**. Esta es una parte fundamental de cualquier ensayo clínico, por lo que debe estar claramente definido cómo se hizo. Aquí os tengo que decir que el azar es caprichoso y tiene muchos vicios, pero raramente produce grupos de igual tamaño. Pensad un momento si tiráis una moneda 100 veces. Aunque la probabilidad de salir cara en cada lanzamiento sea del 50%, será muy raro que lanzando 100 veces saquéis exactamente 50 caras. Cuanto mayor sea el número de participantes, más sospechoso nos deberá parecer que los dos grupos sean iguales. Pero cuidado, esto solo vale para la aleatorización simple. Existen métodos de aleatorización en los que los grupos sí pueden quedar más equilibrados.

Otro punto caliente es el uso indebido que, a veces, puede hacerse con **variables cualitativas**. Aunque las variables cualitativas pueden codificarse con números, mucho cuidado con hacer operaciones aritméticas con ellos. Probablemente no tendrán ningún sentido. Otra trampa que podemos encontrarnos tiene que ver con el hecho de categorizar una variable continua. Pasar una variable continua a cualitativa suele llevar aparejada pérdida de información, así que debe tener un significado clínico claro. En caso contrario, podemos sospechar que la razón sea la búsqueda de una $p < 0,05$, siempre más fácil de conseguir con la variable cualitativa.

Entrando ya en el análisis de los datos, hay que comprobar que los autores han **seguido el protocolo del estudio** diseñado a priori. Desconfiad siempre de los estudios *post hoc* que no estaban planificados desde el comienzo. Si buscamos lo suficiente, siempre hallaremos un grupo que se comporta como a nosotros nos interesa. Como suele decirse, si torturas los datos lo suficiente, acabarán por confesar.

Otra conducta inaceptable es finalizar el estudio antes de tiempo por obtenerse buenos resultados. Una vez más, si la duración del seguimiento se ha

establecido durante la fase de diseño como la idónea para detectar el efecto, esto debe respetarse. Cualquier violación del protocolo debe estar más que justificada. Lógicamente, es lógico terminar el estudio antes de tiempo por motivos de seguridad de los participantes, pero habrá que tener en cuenta cómo afecta este hecho en la valoración de los resultados.

Antes de realizar el análisis de los resultados, los autores de cualquier trabajo tienen que depurar sus datos, revisando la calidad y la integridad de los valores recogidos. En este sentido, uno de los aspectos a los que hay que prestar atención es al **manejo de los datos extremos** (los llamados *outliers*). Estos son los valores que se alejan mucho de los valores centrales de la distribución. En muchas ocasiones pueden deberse a errores en el cálculo, medición o transcripción del valor de la variable, pero también pueden ser valores reales que se deban a la especial idiosincrasia de la variable. El problema es que existe una tendencia a eliminarlos del análisis aun cuando no haya seguridad de que se deban a algún error. Lo correcto es tenerlos en cuenta al hacer el análisis y utilizar, si es necesario, métodos estadísticos robustos que permitan ajustar estas desviaciones.

Finalmente, el aspecto que nos puede costar más a los no muy expertos en estadística es saber si se ha **empleado el método estadístico correcto**. Un error frecuente es el empleo de pruebas paramétricas sin comprobar previamente si se cumplen los requisitos necesarios. Esto puede hacerse por ignorancia o para obtener la significación estadística, ya que las pruebas paramétricas son menos exigentes en este sentido. Para entendernos, la p será más pequeña que si empleamos la prueba equivalente no paramétrica.

También, con cierta frecuencia, se obvian otros requisitos para poder aplicar determinada prueba de contraste. Como ejemplo, para realizar una prueba de la t de Student o un ANOVA es necesario comprobar la homocedasticidad (una palabra muy fea que quiere decir que las varianzas son iguales), comprobación que se pasa por alto en muchos trabajos. Lo mismo ocurre con los modelos de regresión que, con frecuencia, no se acompañan del preceptivo diagnóstico del modelo que permite justificar su uso.

Otro asunto en el que puede haber trampa es el de las comparaciones múltiples. Por ejemplo, cuando el ANOVA da significativo nos dice que hay al menos dos medias que son diferentes, pero no cuáles, así que nos ponemos a compararlas dos a dos. El problema es que cuando hacemos comparaciones repetidas aumenta la probabilidad de error de tipo I, o sea, la probabilidad de encontrar diferencias significativas solo por azar. Esto puede permitir encontrar, aunque solo sea por casualidad, una $p < 0,05$, lo que viste mucho el estudio (sobre todo si has gastado mucho tiempo y/o dinero en hacerlo). En estos casos los autores deben emplear alguna de las correcciones disponibles (como la de Bonferroni, una de las más sencillas) para que el alfa global se mantenga en 0,05. El precio a pagar es sencillo: la p tiene que ser mucho más pequeña para ser significativa. Cuando veamos comparaciones múltiples sin corrección solo tendrá dos explicaciones: la ignorancia del que haya hecho el análisis o el intento de encontrar una significación que, probablemente, no soportaría la disminución del valor de p que conllevaría la corrección.

Otra víctima frecuente del mal uso de la estadística es el coeficiente de correlación de Pearson, que se utiliza para casi todo. La correlación, como tal,

nos dice si dos variables están relacionadas, pero no nos dice nada sobre la causalidad de una variable para la producción de la otra. Otro mal uso es utilizar el coeficiente de correlación para comparar los resultados obtenidos por dos observadores, cuando probablemente lo que deba utilizarse en este caso es el coeficiente de correlación intraclase (para variables continuas) o el índice kappa (para cualitativas dicotómicas). Por último, también es incorrecto comparar dos métodos de medición (por ejemplo, glucemia capilar y venosa) mediante correlación o regresión lineal. Para estos casos lo correcto sería usar la regresión de Passing y Bablok.

Otra situación en la que una mente paranoica como la mía sospecharía es aquella en la que el método estadístico empleado no lo conocen ni los más listos del lugar. Siempre que haya una forma más conocida (y muchas veces más sencilla) de hacer el análisis, deberemos preguntarnos **por qué han usado un método tan raro**. En estos casos exigiremos a los autores que justifiquen su elección y que aporten una cita donde podamos revisar el método. En estadística hay que tratar de elegir la técnica correcta para cada ocasión y no aquella que nos proporcione el resultado más apetecible.

En cualquiera de los test de contraste anteriores, los autores suelen emplear un nivel de significación para $p < 0,05$, lo habitual, pero el contraste puede hacerse con una o con dos colas. Cuando hacemos un ensayo para probar un nuevo fármaco, lo que esperamos es que funcione mejor que el placebo o el fármaco con el que lo estemos comparando. Sin embargo, pueden ocurrir otras dos situaciones que no podemos desdeñar: que funcione igual o, incluso, que funcione peor. Un contraste bilateral (con dos colas) no asume la dirección del efecto, ya que calcula la probabilidad de obtener una diferencia igual o mayor que la observada, en las dos direcciones. Si el investigador está muy seguro de la dirección del efecto puede hacer un contraste unilateral (con una cola), midiendo la probabilidad del resultado en la dirección considerada. El problema es cuando lo hace por otra razón: la p del contraste bilateral es el doble de grande que la del unilateral, por lo que será más fácil conseguir significación estadística con el contraste unilateral. Lo que no es correcto es que este último sea el motivo para hacer un contraste unilateral. Lo correcto, salvo que haya razones bien justificadas, es hacer un contraste bilateral.

Para ir terminando esta entrada tan tramposa, diremos unas palabras sobre el uso de las medidas adecuadas para presentar los resultados. Hay muchas formas de maquillar la verdad sin llegar a mentir y, aunque en el fondo todas dicen lo mismo, la apariencia puede ser muy diferente según cómo lo digamos. El ejemplo más típico es el de usar medidas de riesgo relativas en lugar de medidas absolutas de impacto. Siempre que veamos un ensayo clínico, debemos exigir que nos presenten la reducción absoluta del riesgo y el número necesario a tratar (NNT). La reducción relativa del riesgo es un número mayor que la absoluta, por lo que parecerá que el impacto es mayor. Dado que las medidas absolutas son más fáciles de calcular y se obtienen de los mismos datos que la relativas, deberemos desconfiar si los autores no nos las ofrecen: quizás el efecto no sea tan importante como nos pretenden hacer ver.

Otro ejemplo es el uso de la odds ratio frente a los riesgos relativos (cuando pueden calcularse ambos). La odds ratio tiende a magnificar la asociación entre las variables, así que su uso no justificado también puede hacernos

sospechar. Si podéis, calcular el riesgo relativo y comparad las dos medidas.

De igual manera, sospecharemos de los estudios de pruebas diagnósticas que no nos proporcionan los cocientes de probabilidad y se limiten a sensibilidad, especificidad y valores predictivos. Los valores predictivos pueden ser altos si la prevalencia de la enfermedad en la población del estudio es alta, pero no sería aplicables a poblaciones con menos proporción de enfermos. Esto se soslaya con el uso de los cocientes de probabilidad. Siempre deberemos preguntarnos el motivo que puedan tener los autores para obviar el dato parámetro más válido para calibrar la potencia de la prueba diagnóstica.

Y, por último, mucho cuidado con los gráficos: aquí las posibilidades de maquillar los resultados solo están limitadas por la imaginación. Hay que fijarse en las unidades empleadas y tratar de extraer la información del gráfico más allá de lo que pueda parecer que representa a primera vista.

Y aquí dejamos el tema por hoy. Nos ha faltado hablar en detalle sobre otra de las entidades más incomprendidas y manipuladas, que no es otra que nuestra p. A p se le atribuyen muchos significados, generalmente de forma errónea, como la probabilidad de que la hipótesis nula sea cierta, probabilidad que tiene su método específico para poder hacer una estimación. Pero esa es otra historia…

Bibliografía

1. Altman DG. Poor-quality medical research: what can journal do? JAMA.2002;287:2765-7. (PubMed)

2. Molina Arias M. Razones para dejar de leer un artículo. Rev Pediatr Aten Primaria.2014;16:87-91. (HTML)

3. Molina Arias M. Las trampas de la estadística. Rev Pediatr Aten Primaria.2014;16:181-6. (HTML)

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología.
Hospital Infantil Universitario La Paz.
Madrid. España.

Aceptado para el blog en agosto de 2019