Bioestadística y Epidemiología

DATOS FALTANTES (MISSING VALUES)

Jorge Dagnino S.[1]

- Los datos faltantes son un problema frecuente en los estudios médicos; habitualmente no son reportados y, si se menciona el hecho, no se explicita la manera en que fueron enfrentados.
- Los programas computacionales los manejan en forma variable, lo que puede conducir a errores en los resultados y su interpretación.
- No hay ninguna forma totalmente satisfactoria para el manejo de los datos faltantes, por lo que se debe ser estricto en optimizar la recolección y registro de datos en la etapa de diseño y ejecución.
- El sesgo que introduce o puede introducir la falta de datos es proporcional al número de pérdidas: más de un 10% es inaceptable.
- Alternativas para mitigar los datos faltantes: 1) Omitir variables con datos faltantes. 2) Omitir individuos en quienes hay datos faltantes. 3) Estimar (imputar) los datos faltantes donde estos son reemplazados con valores predichos desde los datos presentes.

Los datos faltantes se definen como valores no disponibles que serían útiles o significativos para el análisis de los resultados. Hay muchos tipos de datos faltantes y muchas razones por las cuales pueden ocurrir. Estos dos factores son decisivos al enfrentar la ausencia de datos en el momento de analizar los resultados, donde lo principal es decidir si la pérdida es aleatoria, es decir, afecta por igual a todos los individuos, o bien puede ser debida a una razón o razones específicas que pueden introducir sesgos que invaliden los resultados.

Es un problema muy frecuente en estudios médicos; no todos lo reportan y menos aún comentan los métodos usados, si alguno, para enfrentarlo. En ensayos clínicos aleatorios, la pérdida de datos diluye la aleatorización, introduce sesgos desconocidos y se compromete la potencia del estudio. Esto representa un grave problema para la confiabilidad de los resultados, donde los datos faltantes pueden estar relacionados con la efectividad de un tratamiento, los efectos adversos o el pronóstico. A pesar de ello, muchos textos no se refieren al problema y muchos programas computacionales asumen que los datos están completos. En 2012, el National Research Council de EE.UU. convocó a un comité para el estudio y proposiciones sobre los datos faltantes.

La manera en que los programas computacionales manejan los datos faltantes no es uniforme y pueden introducir errores mayores en los resultados y conclusiones. Ello se complica más aún si se adoptan diversas maneras de codificar los datos faltantes; si la estrategia incluye usar números reales para señalar un dato faltante el programa puede incluirlos en los cálculos distorsionando en mayor o menor grado los resultados. Por otro lado, muchos programas computacionales simplemente omiten del análisis aquellos individuos que no tienen los datos completos. Esto reduce el tamaño muestral y los autores pueden no detectarlo pues el programa igual termina haciendo cálculos y arrojando una p al final. En un ANOVA, por ejemplo, esta merma sólo se detecta en la reducción de los grados de libertad en la tabla de resultados; otros programas comunican la reducción con una nota al pie pero aquellos autores que sólo están interesados en el valor de p pueden no advertirlo. Si en el manuscrito no aparece completo el análisis del ANOVA, con los grados de libertad usados en el cálculo de F, los lectores no tienen manera de detectar el error.

El problema es más frecuente en trabajos retrospectivos, especialmente cuando los datos recolectados rutinariamente son usados *a posteriori* con otros objetivos. Sin embargo, los datos faltantes son particularmente sensibles en estudios clínicos longitudinales en donde los resultados pueden determinar la presencia y extensión de los datos fal-

[1] Profesor Titular
División de Anestesiología. Pontificia Universidad Católica de Chile.

tantes; por ejemplo, la presencia de efectos adversos puede condicionar el abandono de los pacientes o la pérdida de seguimiento puede ser mayor en pacientes que fallecen. No hay ninguna manera completamente satisfactoria para manejar los datos faltantes por lo que se debe poner énfasis en optimizar la recolección y registro de los datos en la etapa de diseño y durante la ejecución. En la etapa de diseño se proponen diversas estrategias como por ejemplo, diseñar los tratamientos (intervenciones) que tengan flexibilidad como para acomodar diferentes preferencias, hacer el seguimiento lo más corto posible, evitar mediciones que por experiencias previas tienen mayor probabilidad de faltar. En la etapa de ejecución: poner metas y monitorizar frecuentemente los datos faltantes, poner incentivos para los participantes junto con regulaciones éticas estrictas, limitar las cargas o dificultades en la recolección de los datos, ofrecer entrenamiento a los participantes y facilitar los métodos de registro. Estas consideraciones deben estar definidas en el protocolo de los trabajos y descritas en el trabajo publicado.

Al enfrentar el problema de datos faltantes, el punto más importante es decidir si estos pueden introducir sesgos en el análisis. Si no se sabe nada sobre la o las causas por las cuales faltan datos es imposible descartar un posible sesgo y menos estimar su magnitud. Un segundo punto importante es la cantidad: si son pocos los datos faltantes, es probable que su efecto sea menor pero si son muchos su ausencia va comprometiendo progresivamente la validez de las conclusiones. Al respecto no hay una cifra mágica pero probablemente pérdidas mayores al 10% no son aceptables en la mayoría de las circunstancias.

Principios para hacer inferencias sobre los datos faltantes:

- Definir, si es factible, si los datos faltantes son significativos para el análisis (y por ende cumplen con la definición para ser datos faltantes).
- Esto implica definir una medición de un posible efecto causal.
- Documentar en lo posible la o las razones porqué falta cada dato.
- Decidir presunciones principales sobre el mecanismo de datos faltantes siguiendo la clasificación detallada más abajo. Este raciocinio debe estar explícito para los lectores.
- Análisis basado en las presunciones anteriores.
- Evaluar la robustez de estas presunciones.

Los datos faltantes se clasifican en tres categorías usando una terminología que puede confundir pero que está relacionada con la manera en que se aconseja o se puede enfrentar su ausencia.

1) Completamente al azar: el hecho que falte una observación no está relacionado con el o los valores faltantes ni con los valores existentes. Otra manera de pensarlo es que cualquier valor tiene la misma probabilidad de faltar que cualquier otro. Por ejemplo, fallas ocasionales de equipos que impiden hacer una medición, olvido ocasional en registrar un dato, el encargado de hacer la medición se enfermó o pérdidas de muestras porque se rompieron los tubos. Omitir del análisis a los individuos con datos faltantes no alteraría la validez pero podría disminuir la potencia del estudio. Estimar *a priori* posibles pérdidas por este mecanismo debiera formar parte del protocolo en el cálculo del tamaño muestral.
2) Al azar: una o varias características registradas pueden explicar la distribución de los datos faltantes. Por ejemplo: el nivel de respuestas faltantes en una encuesta está relacionado con el nivel socio-económico, el número de pacientes con un ECG preoperatorio está relacionado con la edad de los pacientes, o un centro en un estudio multicéntrico no mide una variable particular porque no cuenta con los medios para ello. El nombre es confundente por lo que algunos prefieren "falta ignorable o manejable" estadísticamente.
3) No al azar: los datos faltantes probablemente dependen o están relacionados con datos no observados. Por ejemplo: falta de respuesta en un cuestionario, pérdida durante el seguimiento. El sesgo o los sesgos que pueden introducirse son evidentes e invalidan en mayor o menor medida los resultados.

Hay tres alternativas para lidiar con los datos faltantes: 1) Omitir variables con datos faltantes; 2) Omitir individuos en quienes hay datos faltantes. Estos dos métodos son los que se usan probablemente con mayor frecuencia pero, como producen una pérdida de información y potencia del estudio y, además, no modifican el riesgo de sesgos, no debieran ser usados; 3) Estimar (imputar) los datos faltantes donde estos son reemplazados con valores predichos desde los datos presentes. La imputación puede ser simple (por ejemplo, usar el último valor registrado, el basal o promedios) o a través de ecuaciones o modelos para calcular los valores faltantes (por ejemplo, asumir que una determinada variable tiene una distribución normal con una determinada media y varianza). Toda imputación basada en modelos se basa en presunciones no verificables por lo que no hay ningún método o modelo generalmente recomendable ni completamente satisfactorio. Sin

embargo, se estima que estos métodos son preferibles a la omisión de casos o la imputación simple.

Cualesquiera sea el método usado debiera hacerse un análisis de sensibilidad para evaluar si las conclusiones pueden variar si las presunciones sobre los datos faltantes cambian. Así por ejemplo, en un estudio de mortalidad donde hay un cierto número de pacientes perdidos al seguimiento, se hace un análisis imputando el mejor escenario, todos los faltantes siguen vivos; en seguida se repite el análisis usando el peor escenario y se comparan las conclusiones.

Una lista de verificación para que el lector crítico pueda estimar si hay o no sesgos importantes o si no se pueden eliminar del todo:

1) Todo estudio debiera mencionar si hubo o no datos faltantes.
2) Informar su magnitud. Mientras más grande sea el estudio, mayor el número de variables medidas, más largo en el tiempo, mayor número de participantes, individuos o instituciones, mayor es la probabilidad de tener datos faltantes.
3) Explicar las razones por las cuales faltan los datos y determinar el tipo de datos faltantes.
4) Explicar el método usado para lidiar con el problema: omitir variables, casos o imputación.
5) Explicar el raciocinio, las presunciones del modelo usado para la imputación y sus posibles sesgos.
6) Explicar si se hizo o no un análisis de sensibilidad.

REFERENCIAS

1. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991.
2. Altman DG, Bland JM. Missing data. BMJ 2007; 334: 424.
3. Armitage P, Berry G. Estadística para la investigación biomédica. 3a ed. Barcelona: Harcourt Brace, 1997.
4. Bland M. An Introduction to Medical Statistics. 3rd Ed, Oxford: OUP, 2006.
5. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59: 1087-1091.
6. Hogan JW, Roy J, Korkontzelou C. Handling drop-out in longitudinal studies. Statist Med 2004; 23: 1455-1497.
7. Little RJ, Cohen ML, Dickersin K, et al. The design and conduct of clinical trials to limit missing data. Statist Med 2012; 31: 3433-3443.
8. Little RJ, D'Agostino RB, Cohen ML, et al. The Prevention and Treatment of Missing Data in Clinical Trials. NEJM 2012; 367: 1355-1360.
9. Ware JH, Harrington D, Hunter DJ, D'Agostino RB. Missing Data. NEJM 2012; 367:1353-1354.

Correspondencia a:
Dr. Jorge Dagnino S.
jdagnino@med.puc.cl