Metodología

Los datos perdidos en estudios de investigación ¿son realmente datos perdidos?

*Dr. Pablo Durán**

La pérdida de datos en investigación es siempre una realidad y un problema a considerar

Es inevitable en todo estudio de investigación, independientemente de su diseño metodológico, la pérdida o no disponibilidad de una proporción variable de los datos correspondientes a los sujetos seleccionados.

Estos datos faltantes pueden involucrar desde algunas de las variables de algunos de los sujetos seleccionados hasta la totalidad de los datos de algunos de los individuos seleccionados.

En estudios retrospectivos basados en registros preexistentes (por ejemplo, historias clínicas) es común que en varios registros no conste alguno de los datos requeridos. El desconocimiento o el no recuerdo por parte del respondente de algunos de los datos requeridos constituye otro de los motivos que frecuentemente lleva a la pérdida de datos. En otros casos no es posible obtener algunos datos, por ejemplo, resultados de determinaciones bioquímicas, por no contar con el consentimiento del paciente para la toma de la muestra de sangre pero sí para responder al cuestionario o por dificultades técnicas (muestra insuficiente, contaminación de cultivo, etc.).

Finalmente, en aquellos diseños muestrales con selección nominal que no permiten reemplazos, la negativa a participar implica la pérdida de la totalidad de los datos.

Los datos perdidos, según sus características y proporción, pueden afectar en forma importante tanto la precisión como la validez de las estimaciones a alcanzar.

¿En qué medida los datos perdidos pueden afectar los resultados?

La precisión se ve afectada por la reducción del número total de casos o de una o más variables, en tanto que la validez se verá afectada debido a los posibles sesgos que pueden determinar la pérdida de valores en sujetos con características comunes (por ejemplo, debido a una mayor frecuencia de negativas en el grupo de condiciones socioeconómicas más elevadas).

El tratamiento de posibles datos perdidos, con el fin de reducirlos o caracterizarlos, debe considerarse fundamentalmente en tres momentos del desarrollo de todo estudio de investigación: a) durante la etapa de diseño, mediante la adición al tamaño muestral mínimo requerido de una proporción variable de sujetos que permita compensar las posibles pérdidas de datos o sujetos, evitando de este modo el no alcanzar el tamaño muestral requerido; b) durante la etapa de recolección de datos, mediante un adecuado monitoreo de la calidad de los datos que permita recuperar datos perdidos. El tercer momento, una vez recolectados los datos, si bien no permite reducir el impacto sobre la precisión y validez, permitirá valorar la medida en que ambos se ven afectados por la pérdida de datos. Este momento corresponde al proceso de análisis de consistencia de los datos y valores perdidos, que debe formar parte de todo estudio, antes de la estimación de los estadísticos que permitan dar respuesta a los objetivos definidos.

¿Cómo valorar esta situación?

El análisis de datos perdidos implica principalmente valorar su proporción y características para las variables de resultado principales y la posible presencia de sesgos en la distribución de valores perdidos.

Si bien algunos paquetes estadísticos cuentan con módulos que permiten analizar valores perdidos a partir de series de datos, la metodología básica no requiere de procesamientos sofisticados.

* Comité Editorial de Archivos Argentinos de Pediatría.

Con el fin de ejemplificar las características de la valoración de datos perdidos, se seleccionó aleatoriamente una muestra de registros a partir de una base de datos del Sistema Informático Perinatal. La muestra incluyó 1.000 registros y las variables incluidas fueron edad materna, edad gestacional en la primera consulta prenatal, edad gestacional al momento del parto, ganancia de peso materno durante la gestación, peso al nacer, índice de masa corporal (IMC) pregestacional y años de educación materna.

A partir de la base original se conformó una segunda base de datos eliminando algunos de los valores de las diferentes variables. La *Tabla 1* presenta, para cada una de las variables mencionadas, el porcentaje de datos perdidos y el número de valores extremos. Este primer aspecto permite valorar que, aun cuando el número total de registros es de 1.000, el número de registros con datos varía entre 996 (edad materna) y 739 (edad gestacional en la primera consulta). La valoración de valores extremos es tan importante como la de datos perdidos, ya que funcionalmente se comportan como datos perdidos. El análisis de valores extremos excede los alcances del presente trabajo; sin embargo, es relevante tener en cuenta que una vez identificados, debe valorarse la necesidad de que sean excluidos del análisis de los datos, perdiéndose, por lo tanto, tales datos.

Una vez cuantificados los datos perdidos, es necesario valorar en quiénes se presentan los datos faltantes. En el ejemplo antes mencionado se observó que en quienes el dato sobre peso al nacer estaba disponible, la frecuencia de bajo nivel de instrucción fue de 13,2%, en tanto que fue de 56,3% en el grupo con dato faltante. Esta observación pone de manifiesto que en términos del nivel de instrucción materna, los registros con datos de peso al nacer están sesgados con respecto al total de los casos.

La identificación de la cantidad y características de los casos con datos perdidos es importante, pero muchas veces no es suficiente para mejorar la precisión o validez de las estimaciones. Existen diferentes metodologías que permiten reemplazar matemáticamente los datos perdidos por valores calculados a partir del resto de los valores como, por ejemplo, mediante regresión o mediante el método EM (expectación-maximización).

En la *Tabla 2* se comparan valores medios y frecuencia de peso al nacer y bajo nivel de instrucción materna en la base de datos completa, la base con datos perdidos y aquellas con valores estimados mediante dos métodos.

Si bien la estimación de valores perdidos puede resultar un procedimiento sencillo si se cuenta con el programa estadístico adecuado, la elección de los procedimientos para el manejo de datos incompletos constituye una tarea compleja. La precisión de las estimaciones varía según el método utilizado y según la distribución y características de los valores perdidos. Las limitaciones en la imputación de datos han sido ampliamente consideradas en la bibliografía; en muchos casos es preferible no realizar imputaciones o bien, en diseños longitudinales en los que se cuenta con múltiples valoraciones en un

TABLA 1. *Número de casos con datos completos, con dato faltante y valores extremos de variables seleccionadas a partir de datos simulados*

	Casos con dato	Perdidos		Nº de datos extremos	
		n	%	Bajos	Altos
Edad materna	996	4	0,4	0	6
Edad gestacional al parto	964	36	3,6	74	24
Edad gestacional en la 1ª consulta	739	261	26,1	0	0
Ganancia de peso	827	173	17,3	1	3
Peso al nacer	916	84	8,4	14	5
IMC pregestacional	745	255	25,5	1	25
Educación materna	936	64	6,4	0	4

TABLA 2. *Comparación entre estimadores obtenidos mediante diferentes métodos de imputación*

	Base completa (n= 1.000)	Base con datos perdidos sin estimación	Base con datos perdidos y estimación por regresión	Base con datos perdidos y por estimación EM
Peso al nacer (g)	3.294	3.291	3.302	3.295
Media (IC 95%)	(3.264-3.323)	(3.260-3.321)	(3.272-3.331)	(3.267-3.324)
Bajo nivel de instrucción materna %	16,9	13,2	16,5	15,8

mismo individuo, la imputación a partir de otros datos del mismo individuo se valora como la forma más adecuada.

Independientemente del camino a tomar para su resolución, todo trabajo debe cuantificar y caracterizar los casos con datos perdidos

En resumen, antes del tratamiento de los datos es imprescindible valorar la proporción y características de los datos perdidos, además de presentar en todo informe científico una descripción detallada del total de casos seleccionados y las pérdidas registradas en relación con el total de casos, así como para cada una de las variables que se presentan. Las imputaciones pueden realizarse, pero considerando que pueden igualmente conducir a estimaciones sesgadas. La imputación a partir de datos de los propios individuos en momentos diferentes o métodos de imputación múltiple constituyen métodos que brindan estimaciones más razonables. Sin embargo, es necesario ser cauteloso en el momento de decidir su implementación. ■

BIBLIOGRAFÍA CONSULTADA

- Rothman KJ, Greenland S. Modern Epidemiology. Philadelphia: Lippincott-Raven Publishers, 1998.
- SPSS Missing Value Analysis. Chicago, IL: 7.5. SPSS Inc., 1997.
- Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol 2003; 56:968-76.

Hace 50 años en Archivos Argentinos de Pediatría

Tratamiento de la tos convulsiva. Nuestra experiencia en 461 casos

Dres. Leonidas Taubenslag, Ilda Moreno de Taubenslag y Tomás Lassalle

(...) **La Higiene mental del niño coqueluchoso** *(Dra. Moreno)*

Un último planteo no despreciable dentro del programa general del tratamiento de la coqueluche quedaría por hacer, y es el que se refiere a la participación que la enfermedad tiene en el psiquismo del niño. Si bien es cierto que cualquier noxa influye en el curso naturalmente plácido de su vida, la coqueluche por lo que tienen sus crisis de aparatoso lo coloca en un estado de ansiedad cuya duración puede ser modificada por la actitud del medio ambiente.

Es de observación frecuente; lo que se ha dado en llamar el automatismo de las quintas, es decir, persistencia del carácter coqueluchoso de una tos ya estéril. Se ha dado diversas explicaciones a este fenómeno, pero es evidente que muchos niños, habiendo captado la influencia que sus crisis de tos tiene como desencadenante de la angustia hogareña, la siguen utilizando como embajadora de sus exigencias desmedidas.

La terapéutica se enriquecerá, entonces, con oportunas y tan frecuentemente olvidadas indicaciones a los padres respecto a la utilidad de la actitud serena y animosa durante las crisis, y a la necesidad de ubicar al niño momentáneamente desconectado de sus actividad normal en un ambiente grato, confortable, (...).

Archivos Argentinos de Pediatría. Año XXVI Noviembre-Diciembre de 1955.
Tomo XLIV, N°. 5, pág. 246-253.