La calidad en la vinculación de datos necesita más atención

Las técnicas de vinculación de datos permiten la identificación y vinculación de datos sobre un mismo individuo almacenados en diferentes bases [1]. Son muchas las posibilidades de aplicación de estas técnicas en la investigación, vigilancia y evaluación en salud, lo que ha suscitado un creciente interés en su utilización. Siguiendo la tendencia internacional, observamos un crecimiento en la remisión de artículos a CSP que emplean técnicas de vinculación de datos. No obstante, son raros los estudios que relatan la calidad de este proceso.

La calidad en la clasificación de enlaces por pares verdaderos o falsos es uno de los aspectos que debe ser evaluado y relatado en los artículos. El proceso de vinculación puede erróneamente clasificar un enlace como un correspondiente verdadero, cuando los registros no pertenecen, de hecho, al mismo individuo (falso positivo), así como dejar de clasificar como correspondiente verdadero un enlace cuyos registros pertenezcan al mismo individuo (falso negativo). Los errores falso positivo se producen más frecuentemente cuando están disponibles pocos campos para la comparación, los campos de identificación presentan baja completitud, la proporción de homónimos es elevada y las bases relacionadas presentan un gran volumen de datos. En el caso de los errores falso negativo, se producen debido a que la información obtenida es incorrecta, a errores de mecanografía y a la ausencia del registro de eventos en las bases. Los errores de vinculación tienen como resultado una mala clasificación de la exposición, de las conclusiones, o de ambos. Esos errores pueden introducir sesgos en la estimativa de las medidas de asociación, especialmente en situaciones donde se produzca dependencia en la mala clasificación de la exposición y desenlace, y cuando los errores son diferenciales [2].

El mayor desafío para la evaluación de la calidad de procesos de vinculación es la disponibilidad de un patrón fiable. Una alternativa, aunque imperfecta, es el empleo de una muestra de enlaces cuyo status está determinado por una revisión manual [1]. En este caso, la muestra debe ser seleccionada de forma que represente todo el conjunto de enlaces formados en el proceso automático. Otra alternativa sería la utilización de conjuntos de datos desarrollados para test [1]. Es necesario el desarrollo de conjuntos de datos para test que representen las bases de salud brasileñas.

Recientemente, la importancia de un mayor rigor y transparencia están siendo enfatizadas en la dirección y elaboración de estudios [3,4]. En este sentido, fueron elaboradas dos directrices orientadas a estudios que emplean técnicas de vinculación de datos [5,6]. Recomendamos que los artículos remitidos a CSP sigan las orientaciones indicadas en esas directrices.

Cláudia Medina Coeli
Editora

1. Christen P. Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection. Heidelberg: Springer; 2012.
2. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. Heidelberg: Springer; 2009.
3. Kac G, Hirst A. Enhanced quality and transparency of health research reporting can lead to improvements in public health policy decision-making: help from the EQUATOR Network. Cad Saúde Pública 2011; 27:1872-3.
4. McNutt M. Journals unite for reproducibility. Science 2014; 346:679.
5. Bohensky MA, Jolley D, Sundararajan V, Evans S, Ibrahim J, Brand C. Development and validation of reporting guidelines for studies involving data linkage. Aust N Z J Public Health 2011; 35:486-9.
6. Dusetzina SB, Tyree S, Meyer A-M, Meyer A, Green L, Carpenter WR. Linking data for health services research: a framework and instructional guide. Rockville: Agency for Healthcare Research and Quality; 2014.