We must pay more attention to record linkage quality

Record linkage techniques enable researchers to identify and merge data regarding a single individual stored in different databases [1]. There are many possibilities for applying these techniques in health research, surveillance and evaluation. This has led to an increasing interest in their use. Following an international trend, we have observed a growth in the number of articles submitted to CSP that employ record linkage techniques. However, only a few of these studies report the quality of the linkage process.

One of the aspects that need to be evaluated and reported in the articles is the quality of the classification of links in true or false matches. The linkage process may mistakenly classify a link as a true match when records do not belong to the same individual (false positive), as well as fail to classify a link as a true match when its records do belong do the same person (false negative). False positive errors are more frequent when few fields are available for comparison, completeness of identifiers is low, the proportion of homonyms is high and linked databases have a high volume of data. False negative errors, on the other hand, happen due to incorrect information, typographical errors and the absence of records of the events in the databases. Linkage errors result from misclassification of exposure, outcome or both. They can bias the estimates of association measures, especially in situations in which there is dependence on the misclassification of exposure and outcome and when the errors are differential [2].

The biggest challenge in evaluating the quality of linkage processes is the availability of a gold standard. An alternative, albeit an imperfect one, is to use a sample of links, the status of which is determined by manual revision [1]. In that case, the sample must be selected so as to represent the entire set of links formed by the automatic process. Another alternative is to use sets of data developed for testing [1]. We need to develop test data sets that represent Brazilian health databases.

Recently, different authors have emphasized the need for greater rigor and transparency in conducting and reporting studies [3,4]. Two guidelines for studies employing record linkage techniques have been formulated that address these concerns [5,6]. We recommend that articles submitted to CSP follow these guidelines.

Cláudia Medina Coeli
Editor

http://dx.doi.org/10.1590/0102-311XED010715

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.