A qualidade do *linkage* de dados precisa de mais atenção

Técnicas de *linkage* de dados permitem a identificação e vinculação de dados de um mesmo indivíduo armazenados em diferentes bases [1]. São muitas as possibilidade de aplicação dessas técnicas na pesquisa, vigilância e avaliação em saúde, o que tem suscitado um crescente interesse na sua utilização. Acompanhando a tendência internacional, observamos um crescimento da submissão de artigos a CSP que empregam técnicas de *linkage* de dados. Entretanto, são raros os estudos que relatam a qualidade do processo de *linkage*.

A qualidade da classificação de *links* em pares verdadeiros ou falsos é um dos aspectos que deve ser avaliado e relatado nos artigos. O processo de *linkage* pode erroneamente classificar um *link* como um par verdadeiro, quando os registros não pertencem de fato ao mesmo indivíduo (falso positivo), assim como deixar de classificar como par verdadeiro um *link* cujos registros pertençam ao mesmo indivíduo (falso negativo). Os erros falso positivo ocorrem mais frequentemente quando são disponíveis poucos campos para a comparação, os campos de identificação apresentam baixa completitude, a proporção de homônimos é elevada e as bases relacionadas apresentam grande volume de dados. Já os erros falso negativo ocorrem devido à informação obtida ser incorreta, a erros de digitação e à ausência do registro de eventos nas bases. Erros de *linkage* resultam da má classificação da exposição, do desfecho, ou de ambos. Esses erros podem introduzir viés na estimativa das medidas de associação, especialmente nas situações em que ocorra dependência na má classificação da exposição e desfecho, e quando os erros são diferenciais [2].

O maior desafio para a avaliação da qualidade de processos de *linkage* é a disponibilidade de um padrão ouro. Uma alternativa, ainda que imperfeita, é o emprego de uma amostra de *links* cujo *status* é determinado por revisão manual [1]. Nesse caso, a amostra deve ser selecionada de forma a representar todo o conjunto de *links* formados no processo automático. Outra alternativa seria a utilização de conjuntos de dados desenvolvidos para teste [1]. É necessário o desenvolvimento de conjuntos de dados para teste que representem as bases de saúde brasileiras.

Recentemente, vem sendo enfatizada a importância do maior rigor e transparência na condução e relato de estudos [3,4]. Nesse sentido foram elaborados duas diretrizes orientadas para estudos que empregam técnicas de *linkage* de dados [5,6]. Recomendamos que os artigos submetidos a CSP sigam as orientações indicadas nessas diretrizes.

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.