International Journal of Population Data Science

Journal Website: www.ijpds.org

Assessing the accuracy of probabilistic record linkage of social and health databases in the 100 million Brazilian cohort

Barreto, Marcos[1*], Alves, André[1], Sena, Samila[1], Fiaccone, Rosemeire[1], Amorim, Leila[1], Ichihara, Maria Yuri[1], and Barreto, Mauricio[2]

[1]Federal University of Bahia (UFBA)
[2]Oswaldo Cruz Foundation (FIOCRUZ)

Background and aims

The Brazilian government has several social protection programmes that select their beneficiaries based on socioeconomic information kept in the CadastroÚnico (CADU) database. The CADU will be used to build a population-based cohort of approximately 100 million individuals. Among the social programmes is the Bolsa FamÃŋlia (PBF), a conditional cash transfer programme that provides extra income to poor families. These two databases must be deterministically linked to individuals who have received payments from PBF between 2004 and 2012. It will be used in epidemiological studies aiming to assess the impact of PBF on the occurrence and severity of several diseases and health problems (tuberculosis, leprosy, HIV, child health etc). This cohort must be probabilistically linked with databases from the Unified Health System (SUS), such as hospitalization, notifiable diseases, mortality, and live births, in order to produce data marts (domain-specific data) to the proposed studies. Our goals comprise the validation of probabilistic record linkage methods to support this cohort setup.

Approach

This paper emphasizes the accuracy assessment of our methods based on the linkage of SIH (hospitalization), SINAN (notifications), and SIM (mortality) records to the 2011 extraction of CADU. We focused on hospitalization and notification of tuberculosis, as well infant mortality for all causes in under-4 children, for a small sample with 30,029 records (CADU). Due to the absence of gold standards, we used two approaches to assess accuracy: a clerical review and an automatic (tool-based) search. In the first case, we used different cut-off points as similarity index to calculate sensitivity and specificity, and a ROC curve to separate matched and non-matched pairs. The second approach retrieves from CADU all matched and non-matched pairs for a given individual, serving as a gold standard for validation.

*Corresponding Author:
Email Address: marcosebarreto@gmail.com (M. Barreto)

Results

We retrieved 22 linked pairs, from which 18 are true positives for infant mortality (SIM database). From SINAN, our results were 434 linked pairs with 166 true positives, and with SIH, 121 linked pairs with 34 true positives. The sensitivity of manual scan for SIM (children mortality) ranges from 44% (specificity of 100%) to 95% (specificity of 94%), with similarity indices between 0.80 and 0.97, respectively. For automatic search, we obtained a sensitivity of 69.2% and specificity of 91.8%.

Conclusion

Our results show the need for a continuous improvement in our linkage routines and how to consistently evaluate their accuracy in the absence of adequate gold standards.

http://dx.doi.org/10.23889/ijpds.v1i1.276