Integrating Brazilian health databases

Subjects: Computer Science, Information Systems

Contributor: Patricia Takako Endo, Maicon Herverton Lino Ferreira da Silva Barros

The volume of data generated by health systems is substantial and is likely to continue growing exponentially with the growing adoption of the Internet of Things. Efforts to improve data discovery and integration are complicated by the complexity, dimensionality and heterogeneity of the data, inadequate data, and other data quality issues. This work-in-progress has as its main goal the integration of two Brazilian health databases in order to improve the quality of tuberculosis mortality data. A phonetic encoding technique (Soundex) and a pattern matching recognition (Jaro) are proposed as solutions and results compared. Both techniques identified over 500 true matches with Jaro discovering more true matches than Soundex.

Keywords: e-health ; record linkage ; health database

In Brazil, the Federal Government is responsible for maintaining and updating databases such as the *Sistema de Informação de Agravos de Notificação* (SINAN), the Brazilian National Notifiable Disease Information System. SINAN contains data on patients diagnosed with a notifiable disease and provides access to this data to health professionals across Brazil. Similarly, the *Sistema de Informação sobre Mortalidade* (SIM) provides data on mortality. The databases suffer from major data issues including consistency, completeness, timeliness, and integration.

Despite efforts made by the Ministry of Health to standardize the entire process of collecting, recording and maintaining health data in Brazil, information centralization remains a significant challenge[1]. Data collected by different entities throughout the health system do not have unique identifiers therefore combining these different databases is not a trivial task. These factors, in addition to ICT limitations and the availability of qualified human resources, limit the use of data to generate evidence to support clinical and policy decisions, and to address important epidemiological questions[1] [2] [3].

Despite the efforts made by the Ministry of Health to harmonize the process of collecting, recording and maintaining information in Brazil, there is still a great difficulty in integrating the various existing health institutions in the country. The data collected by different health services do not have unique identifiers, and therefore, combining these different data sources is not a trivial task. These factors, in addition to restrictions on technological infrastructure and qualified human resources, limit the use of data to generate evidence to support clinical and political decisions, and to answer important epidemiological questions.

Against this backdrop, it is essential to apply a record linkage process between the Brazilian health databases. Record linkage (entity resolution or data matching) is "[.] *the process of identifying sets of records that correspond to the same individual*" [4]. In this process, it is common to apply techniques to combine and merge records to identify data that correspond to the same entity in the real world. There are a wide range of techniques for this purpose. Phonetic encoding techniques[5] such as Soundex, Metaphone and Phonex, and pattern matching recognition such as Jaro, Jaro-Winkler, Edit Distance or Levensthein, and Q-grams, are two such techniques [6][4] .

In order to improve data quality and associated decision making, this work-in-progress aims to integrate two Brazilian health databases (SINAN and SIM) using record linkage techniques. The pilot use case is the identification of tuberculosis (TB) deaths in order to create an integrated database containing demographic, clinical and mortality data in relation to TB. This integrated base is of paramount importance for training deep learning models to identify disease severity and predict epidemics.

Preliminary results

We are considering two distinct Brazilian databases: SINAN and SIM, containing data from January 2007 to August 2018 related to the State of Amazonas. The SINAN database has 36,209 records of TB cases, of which 1,224 are fatalities. The SIM database has 205,290 records with 2,866 records related to tuberculosis fatalities. Although the SINAN includes the mortality outcome; it is widely agreed that this data does not reflect the true number of fatality records as many records

are not updated when the patients die. In this way, our main goal is to find patients in SIM that were not recorded as dead in SINAN as we want to create a more accurate database of demographic, clinical and laboratory data by integrating the data between the databases.

The SINAN database had initially 1,224 records of TB fatalities. Following application of record linkages, 534 new records from Soundex were identified for adding to the SINAN database, and 581 new records from Jaro. Currently, the detection of the false matches is a manual process, meaning that if the dataset is larger than this, the task could take longer to complete. The entire discussion of these results can be found in[7].

The entire article has been published on: https://ieeexplore.ieee.org/abstract/document/9139699

References

1. M. S. Ali, M. Y. Ichihara, L. C. Lopes, G. C. Barbosa, R. Pita, R. P.Carreiro, D. B. dos Santos, D. Ramos, N. Bispo, F. Raynalet al.,“Administrative data linkage in brazil: potentials for health technology assessment,”Frontiers in pharmacology, vol. 10, 2019.
2. A. P. Hassler, E. Menasalvas, F. J. Garcia-Garcia, L. Rodriguez-Maas,and A. Holzinger, “Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome,” BMCmedical informatics and decision making, vol. 19, no. 1, p. 33, 2019.
3. R. S. Pinheiro, L. V. Andrade, and G. Oliveira, “Underreporting of tuberculosis in the information system on notifiable diseases (sinan): primary default and case detection from additional data sources using probabilistic record linkage,” Cadernos de saude publica, vol. 28, 2012.
4. C. Nanayakkara, P. Christen, and T. Ranbaduge, “Robust temporal graph clustering for group record linkage,” in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2019, pp. 526–538.
5. M. del Pilar Angeles, A. Espino-Gamez, and J. Gil-Moncada, “Comparison of a modified spanish phonetic, soundex, and phonex coding functions during data matching process,” in2015 International Conference on Informatics, Electronics & Vision (ICIEV). IEEE, 2015.
6. P. Christen, Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & BusinessMedia, 2012.
7. da Silva, M.H.L.F., da Silva Leite, M.T., Sampaio, V., Lynn, T. and Endo, P.T., 2020, June. Application and analysis of record linkage techniques to integrate Brazilian health databases. In 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA) (pp. 1-2). IEEE.