International Journal of Population Data Science

Journal Website: www.ijpds.org

Design and evaluation of probabilistic record linkage methods supporting the Brazilian 100-million cohort initiative.

Pita, Robespierre¹, Pinto, Clicia¹, Barreto, Marcos^{1*}, Sena, Samila¹, Fiaccone, Rosemeire¹, Amorim, Leila¹, and Barreto, Mauricio²

¹Federal University of Bahia (UFBA)

²Oswaldo Cruz Foundation (FIOCRUZ)

Background and aims

A cooperation Brazil-UK was set in mid-2013 aiming at to build a huge cohort comprised by individuals registered in CadastroÚnico (CADU), a socioeconomic database used in social programmes of the Brazilian government. Epidemiologists and statisticians wish to assess the impact of Bolsa Família (PBF), a conditional cash transfer programme, on the incidence of several diseases (tuberculosis, leprosy, HIV etc). The cohort must contain all individuals who received at least one payment from PBF between 2007 and 2012, which results in a 100-million records according to our preliminary analysis. These individuals must be probabilistically linked with databases from the Unified Health System (SUS), such as hospitalization (SIH), notifiable diseases (SINAN), mortality (SIM), live births (SINASC), to produce data marts (domain-specific data) to the proposed studies. Within this cooperation, our first goal was to design and evaluate probabilistic methods to routine link the cohort, PBF, and SUS outcomes.

records) and SINAN (624 records), for tuberculosis in Sergipe, returned 397 (full probabilistic) and 311 (hybrid) linked pairs, being 306 and 300 true positives. Another execution considering CADU (1,988,599 records) and SINAN (2,094 records), for tuberculosis in Santa Catarina, returned 791 (full probabilistic) and 500 (hybrid) linked pairs, with 667 and 472 true positives. Linking CADU (1,685,697 records) and SIM, for mortality of children under-4, returned 18 linked pairs, all of them true positives, for a Dice between 0.90 and 0.92 and with 100% of sensitivity, specificity, and positive predictive value.

Conclusion

Due to the absence of gold standards, we use samples with increasing sizes and manual review when adequate. Our results are quite accurate, although obtained with an unique extraction of CADU. We are starting to run linkages with the entire cohort.

Approach

We implemented two probabilistic linkage methods: a full probabilistic, based on the Dice similarity (Sorensen index) of Bloom filters; and an hybrid approach, based on rules to deterministic and probabilistic matching. We performed linkages involving CADU (2011 extraction) and SUS outcomes (SIH, SINAN, and SIM) with samples from 3 states (Sergipe, Santa Catarina and Bahia) with an increasing size (from 1,447,512 to 12,036,010).

Results

Using a Dice between 0.90 and 0.92, our methods retrieved more than 95% of true positive pairs amongst the linked pairs. For Sergipe, we obtained as <linked pairs,true positives>: <23,22>, <315,300>, <32,32>, respectively for SIH, SINAN, and SIM. For Bahia: <771,593>, <2677,2626>, <208,207>. Another linkage between CADU (1,447,512

*Corresponding Author:

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