Accuracy of probabilistic linkage: the Brazilian 100 million cohort*

Clicia Pinto[1], Robespierre Dantas[1], Samila Sena[2], Sandra Reis[2], Rosemeire Fiaccone[2], Leila Amorim[2], Mauricio Barreto[3], Spiros Denaxas[4,±], Marcos Barreto[1,4], *Member, IEEE*

***Abstract*— We present current results from our probabilistic linkage methods applied to the integration of a 100 million cohort composed by socioeconomic data with health databases.**

I. Introduction

The 100 million cohort project [1] was set up in 2013 aiming at to build a huge population-based cohort to be used by epidemiologists and statisticians to assess the effects of Brazilian social programmes on health and other outcomes.

Due to the nature of the databases involved, we deployed a probabilistic record linkage pipeline that is used to link this cohort with different health databases and produce specific data (*data marts*) to the desired epidemiological studies

II. Data Scope and Methodology

Socioeconomic data are kept in CadastroÚnico (CADU), a central registrar for all social programmes kept by the Brazilian government. Individuals got an unique identifier (NIS) and must renew their information biennially. The cohort has all individuals from CADU who have received payments from Bolsa Família (conditional cash transfer programme) between 2007 and 2015, resulting in a 114 million records so far.

NIS is used to aggregate payment data for each individual. In order to link the cohort with health databases, we use a probabilistic approach due to the absence of common key attributes. We implemented a 4-stage pipeline for data quality analysis, data preparation, record linkage, and accuracy ascertainment. Our Spark-based tool (AtyImo) performs the first three stages.

Quality analysis considers the percentage of missing attributes and other characteristics of the Brazilian health system, serving to choose the best linkage attributes. Data preparation covers standardization, anonymization (through Bloom filters), and blocking. We use a hybrid linkage approach: deterministic comparison applied to some attributes and probabilistic, based on Sørensen–Dice, to others. Linkage is a 2-step process: records are classified as true positives (*match*) or true negatives (*unmatch*), based on specific upper and lower cut-off points. All dubious records pass through a second round, in order to improve accuracy.

* Research supported by FAPESB, Brazilian Research Council (CNPq), Brazilian Ministry of Health, UK Medical Research Council, The Royal Society (UK), Bill & Melinda Gates Foundation.

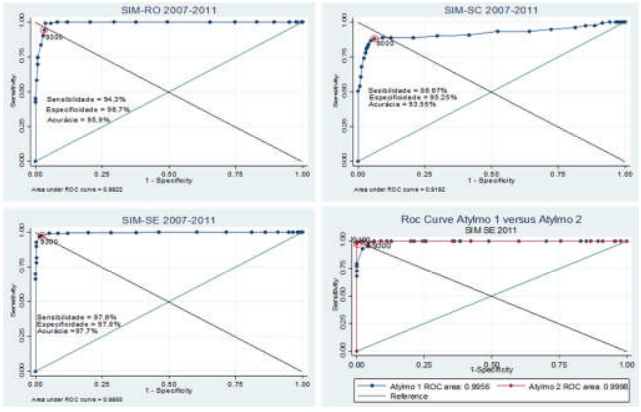
1 Computer Science Dept., Federal University of Bahia (UFBA), Salvador, Brazil.
2 Dept. of Statistics, UFBA, Salvador, Brazil.
3 Oswaldo Cruz Foundation (FIOCRUZ), Salvador, Brazil.
4 Farr Institute of Health Informatics Research, University College London (UK). ± Corresponding author / presenter: s.denaxas@ucl.ac.uk.

III. Accuracy Assessment and Results

Accuracy is measured through sensitivity, specificity and predictive positive value (PPV). We use samples from different federation units (SE, SC and RO) as they have incremental sizes (number of individuals in the cohort) and variable data quality, allowing us to perform manual review of dubious records whenever possible.

Fig. 1 shows the overall cut-off points providing better results to each sample (SE, SC and RO) from a mortality database (SIM) against the cohort, chosen after a year by year (2007 to 2011) analysis. We performed tests with other databases (hospitalizations and notifiable diseases) and calculate sensitivity and PPV for each scenario. In the example, cut-off points vary between 0.90 and 0.93. We also show (bottom right) the accuracy improvement provided by AtyImo v2 (hybrid, 2-step approach) compared to AtyImo v1 [2], which uses a one-step comparison of Bloom filters.

Figure 1. Observed cut-off points and AtyImo's accuracy.



IV. Conclusion

We observed the need of using different cut-off points even considering the same database. Manual review of dubious records is limited by the amount of data to be revised. These issues complicate the definition of gold standards for probabilistic linkage, especially in our 114 million context. We are using machine learning techniques to improve accuracy and trying to eliminate manual review.

References

[1] Mauricio Barreto et al. *Designing and evaluation of probabilistic record linkage methods supporting the Brazilian 100 million cohort initiative*. IPDLN 2016.

[2] Robespierre Pita et al. *A Spark-based workflow for probabilistic record linkage of healthcare data*. BeyondMR Workshop, EDBT/ICDT, 2015.