Assessing the impact of social policies on health in Brazil: the 100 million cohort project

Mauricio Barreto[1,2,#], Maria Yuri Ichihara[2], Rosemeire Fiaccone[3], Leila Amorim[3], Marcos Barreto[4,*], Laura Rodrigues[5,#], Liam Smeeth[5], Katie Harron[5], Spiros Denaxas[6]

This short paper presents our ongoing effort in designing a methodology and associated tools to support longitudinal studies based on a cohort comprised by 100 million individuals. This cohort is used to assess the impact of social programmes held by the Brazilian government, specially *Bolsa Família*, in the incidence of some diseases (tuberculosis, leprosy, HIV) over the beneficiary families.

Socio-economic data from poor families are kept in the CadastroÚnico database. These data are used by the government to select recipients for its social programmes. *Bolsa Família* is one of the largest conditional cash transfer programmes worldwide. Beneficiary families are selected according to their income and receive different amounts of money. In return, they must comply with some conditionalities related to education and health. All payments from *Bolsa Família* are kept in the PBF database and is associated with an individual registered in the CadastroÚnico database.

In Public Health, several databases are used by the Ministry of Health and its agencies to record all related information. Within this project, some databases are specially interesting: SIH (hospitalization), SINAN (notifiable diseases), SIM (mortality), and SINASC (live births). They present different data quality indicators and lack common key attributes.

One key challenge in our project is the volume of data to be correlated. The cohort comprises individuals registered in CadastroÚnico from 2007 to 2012, totalling around 103 million records. These records must be linked with the health databases through a probabilistic approach in order to identify the desired outcomes (diseases) related to beneficiary individuals from *Bolsa Família*.

Based on previous interactions between researchers from Brazil and UK, in 2013 we started this joint effort to develop a methodology and tools to support this project. Our methodology must deal with several technical issues related to data quality and profiling, data conditioning; probabilistic record linkage, and accuracy ascertainment. We designed a prototype implementation based on SPSS, R, and Spark to address these issues. Our results show that our algorithms are able to provide timely executions (from 5 to 9 hours depending on the databases involved in the correlation) and very accurate results (based on sensibility and PPV) in controlled (number of linked pairs can be estimated in advance) and non-controlled (totally probabilistic) scenarios.

We will discuss theoretical and practical aspects in designing this methodology and tool to support the linkage of all these databases, emphasizing: i) how we assess data quality and perform database profiling; ii) the techniques used for data standardization, cleansing, anonymization, and blocking; iii) the implementation of different probabilistic routines for record linkage; iv) how we assess accuracy when gold standards are not available, and v) the data model and database structures we are using to implement the cohort.

1 Oswaldo Cruz Foundation (FioCruz, Bahia), Brazil. 2 Institute of Collective Health (ISC), Federal University of Bahia (UFBA), Brazil. 3. Department of Statistics, Federal University of Bahia (UFBA), Brazil. 4. Computer Science Department, Federal University of Bahia (UFBA), Brazil. 5. London School of Hygiene and Tropical Medicine, London, UK. 6. Farr Institute of Health Informatics Research, London, UK.
* Corresponding author (marcosb@ufba.br). # Principal investigators.