Probabilistic Integration of Large Brazilian Socioeconomic and Clinical Databases

Conference Paper · June 2017

DOI: 10.1109/CBMS.2017.64

9 authors, including:

Marcos Barreto
University College London
46 PUBLICATIONS **236** CITATIONS
SEE PROFILE

Robespierre Pita
Universidade Federal da Bahia
9 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Mauricio L. Barreto
Fundação Oswaldo Cruz
586 PUBLICATIONS **10,956** CITATIONS
SEE PROFILE

Spiros Denaxas
University College London
110 PUBLICATIONS **1,247** CITATIONS
SEE PROFILE

2017 IEEE 30th International Symposium on Computer-Based Medical Systems

Probabilistic integration of large Brazilian socioeconomic and clinical databases

Clicia Pinto*, Robespierre Pita*, George Barbosa*, Bruno Araújo*, Juracy Bertoldo*,
Samila Sena*, Sandra Reis*, Rosemeire Fiaccone*, Leila Amorim*,
Maria Yuri Ichihara*, Mauricio Barreto*, Marcos Barreto*†, Spiros Denaxas†
*Centre for Data and Knowledge Integration for Health (CIDACS), FIOCRUZ, Salvador, Brazil
Email: cliciasp@ufba.br, {pierrepita,gcgbarbosa,arajo.bruno,juracyjuracy,mylasenna}@gmail.com,
{ssreis,fiaccone,leiladen}@ufba.br, {maria.ichihara,mauricio.barreto}@bahia.fiocruz.br
†Farr Institute of Health Informatics Research, University College London (UCL), London, UK
Email: {m.barreto,s.denaxas}@ucl.ac.uk

***Abstract*—The integration of disparate large and heterogeneous socioeconomic and clinical databases is considered essential to capture and model longitudinal and social aspects of diseases. However, such integration is challenging: databases are stored in disparate locations, make use of different identifiers, have variable data quality, record information in bespoke purpose-specific formats and have different levels of metadata. Novel computational methods are required to integrate them and enable their statistical analyses for epidemiological research purposes. In this paper, we describe a probabilistic approach for constructing a very large population-based cohort comprised of 114 million individuals using linkages between clinical databases from the National Health System and administrative databases from governmental social programmes. We present our data integration model for creating data marts (epidemiological data) and discuss our evaluation results in controlled and uncontrolled scenarios, which demonstrate that our model and tools achieve high accuracy (minimum of 91%) in different probabilistic data integration scenarios.**

***Keywords*-Data integration; Probabilistic linkage; Health and social care data; Accuracy assessment.**

I. Introduction

Data integration is a crucial component across several application domains (research, finance, government etc) as it enables the capture and analysis of large volumes of heterogeneous data [1]. In the context of clinical and epidemiological research, data integration arises from the need to combine heterogeneous data sources from diverse sources (hospitals, outpatient clinics, insurance companies, government entities and other administrative sources) to obtain relevant social and health data on study participants.

Epidemiological research heavily relies on this kind of integration to conduct ecological and longitudinal studies based on population samples (cohorts) [2]. The former is characterized by small samples observed over a short period and generally for a specific outcome, whereas the latter utilizes larger samples and observations of several and possibly simultaneous outcomes.

This work pertains to a Brazil-UK cooperation, started in 2013, to provide a computing framework to routinely integrate data from disparate sources (health, education, employment etc) and provide novel analytical methods and tools for researchers to perform data analysis. The primary aim of the project was the creation of a population-based cohort comprised by individuals who have received payments from a conditional cash transfer programme between 2007 and 2015, and its linkage to other health, surveillance and governmental sources for epidemiological research. This resulted in a very large database with 114 million individuals, representing more than 50% of the Brazil population.

Due to the lack of a common and unique person identifier, the integration between administrative and health databases is achieved through probabilistic routines using a set of demographic and person characteristics. Due to the lack of gold standards, the use of probabilistic linkage approaches mandates the design and evaluation of specialized metrics to assess the accuracy of results [3].

In this paper, we describe our approach for probabilistically linking large and heterogeneous health and administrative databases for research. We present our methods to build this huge cohort and address its data heterogeneity. We also discuss how we address data quality assessment and harmonization (transformation, cleansing, anonymization and blocking). Finally, we present some experiments and discuss our accuracy and performance results.

This paper is organizing as follows: Section II presents some related work on record linkage tools and cohort-based initiatives. Section III describes the databases we are using and our approaches to build a huge population-based cohort and implement a record linkage pipeline targeted to integrate this cohort with health databases. Some current results are discussed in Section IV, emphasizing accuracy and scalability. Finally, we present some conclusions and future work in Section V.

II. Related work

In this section, we describe some similar approaches and methods within the wide body of research related to data integration, probabilistic linkage and accuracy assessment.

In the context of clinical research, data integration is used to build cohorts and allow the assessment of policies or

2372-9198/17 $31.00 © 2017 IEEE
DOI 10.1109/CBMS.2017.64

to find data patterns, such as done in the ALSPAC[1] and CONCORD[2] projects, as well in [4] and [5]. Regarding the Brazilian databases we use, there are diverse cohort-based and ecological studies, such as [6] and [7], but they are, in general, based on small samples from specific outcomes (leprosy, malaria, children nutrition etc) linked through traditional database or deterministic linkage tools.

The conventional method for record linkage, based on the pairwise comparison of records from different data sources, was proposed by [8]. It is also widely discussed in [9] and grounded further development [10].

Common data preprocessing methods involved in data linkage, such as cleansing and harmonization, are not widely discussed in literature despite their critical contribution to ensure high accuracy. Doan's book [1] is a good reference for data preparation issues. In [11], the authors conclude that data cleaning can represent up to 75% of the linkage effort. Some proposals to privacy preservation using Bloom filters are presented in [12], [13], and [14]. Blocking and indexing methods are discussed in [10].

There are several tools for probabilistic record linkage currently available, proposed both by the academy and the industry. RecLink [15] was a pioneer proposal targeted to Brazilian databases. German RLC[3], Frill [16] and Febrl [17] are well-known worldwide. CALIBER[4] is a platform integrating EHR (electronic health records) from different databases and supporting a vast range of studies across UK. Dataladder[5] is a tool specifically designed for data cleansing.

Our work differs from existing research in terms of: i) the unprecedented complexity, size, and variability of the health and administrative databases being integrated which contain more than 1 billion rows of data from 114 million participants; ii) the unique set of challenges presented by this task in terms of defining assessment metrics (gold standards), setting reference values (cut-off points) and designing highly-accurate probabilistic linkage routines; and iii) the statistical methods (*Propensity Score Matching, Regression Discontinuity Design, Difference-in-Differences*) intended to be used in the proposed studies, which are feasible to be tested over probabilistic, big data scenarios [3].

III. Proposed approach

The aim of this work was to develop novel probabilistic linkage methods applied to Brazilian governmental databases. More precisely, we needed to i) design a strategy to build a huge population-based cohort aggregating socioeconomic and income transfer data, and ii) implement such methods to link this cohort with health databases and generate "*data marts*" for diverse epidemiological studies.

[1]http://www.bristol.ac.uk/alspac
[2]http://csg.lshtm.ac.uk/research/themes/concord-programme
[3]http://www.record-linkage.de/
[4]https://www.ucl.ac.uk/health-informatics/caliber
[5]https://dataladder.com/

A. Governmental databases

Our methods currently integrate data from six databases: CadastroÚnico (CADU), socioeconomic data (2004-2015); Bolsa Família (PBF), a conditional cash transfer programme (2007-2015); SINASC, birth registry (2001-2012); SIM, mortality registry (2000-2012); SINAN, notifiable diseases (2000-2012); and SIH, hospitalization data (1998-2012).

CADU is a database with socioeconomic data from individuals intending to participate in several social protection programmes. When an individual is registered, an unique and persistent identifier (NIS) is assigned and used to track him across the programmes. Bolsa Família is a well known welfare programme. Individuals registered in CADU and considered poor (according to specific criteria) receive monthly payments and must, in return, comply with a set of conditions. All payments are registered in the PBF database along with the corresponding NIS of each individual.

From a public health perspective, the government has two main strategies to provide free access to health services. Despite being centrally managed, data from these strategies are stored in approximately 40 disparate databases, among which are SIH, SIM, SINAN and SINASC, all of them presenting different structure and data quality.

Common problems associated with these databases include: a) high rates of missing data for specific groups (e.g. homeless people or young children); b) inconsistent coding and recording patterns; c) the absence of an unique, unified, and persistent participant identifier that spans health and administrative datasets. These challenges have significant implications related to the selection of common attributes to probabilistically link these databases.

B. Research cohort setup

The cohort comprise all individuals registered in CADU, between 2007 and 2015, whose received at least one payment (PBF) within this period. To build it, we dealt with three key problems: i) data harmonization between CADU versions; ii) treatment of multiple NIS; and iii) progressive merge of CADU instances.

CADU has two versions: *v6*, from 2007 to 2010, with data organized in two table groups: *Residences* (R, 42 attributes) and *Individuals* (I, 107 attibutes), and *v7*, from 2011 onwards), with 18 tables (additional data on income, work, homeless and disable people, family changes etc) totalizing 433 attributes, from which we used table *1* (*Residence*, 42 attributes) and table *2* (*Individuals*, 38 attributes).

Common attributes to both versions were iteratively evaluated and included in a *inner merge* based on *family_code*, followed by data normalization routines to convert dates and adjust categorical variables. As result, we generated a "baseline" with 15 attributes from each individual: *name, family_code, gender, family_memberID, date_ofBirth, mother_name, code_cityOfBirth, parentage_code, current_NIS,*

Table I: Dimensions of CADU tables

Year	Table	File Size (GB)	Number of records
2007	R	11,4 GB	21.028.364
	I	86.8 GB	79.050.446
2008	R	12.5 GB	22.767.472
	I	100.1GB	89.915.568
2009	R	13.5 GB	24.661.693
	I	108.8 GB	97.640.845
2010	R	14.3 GB	26.107.223
	I	114.4 GB	102.663.287
2011	1	25 GB	27.014.194
	4	4.3 GB	106.433.938
2012	1	11 GB	30.268.867
	4	27 GB	115.636.503
2013	1	6.5 GB	32.897.120
	4	29 GB	123.116.446
2014	1	7.1 GB	35.439.015
	4	34 GB	130.430.300
2015	1	7.6 GB	35.439.015
	4	36 GB	136.368.326

original_NIS, registration_date, registration_status, municipality_code, renewal_date.

Registration in CADU is renewed biannually, which leads to the possibility of individuals holding multiple NIS due to several reasons. They change their family due to marriage or divorce, receiving a new *family_code* that keeps assigned to their NIS. For registration purposes, NIS codes can be active, inactive, blocked or under review, but we retain all NIS regardless of their status. As each CADU instance (year) aggregates data from new and existing individuals, a NIS can be assigned to an individual with different family codes or different NIS are assigned to the same individual.

Our approach to deal with multiple NIS has two phases. Firstly, we use *current_NIS* as a search key to group all records into a "container". Then, we sort this container by *renewal_date* and pick the oldest record to the baseline (as it represents the conditions an individual had before any intervention). The next step is to aggregate all *original_NIS* an individual has into a list to allow us to retrieve all his payments from PBF. At the end, we change *original_NIS* by *LISTOF_original_NIS* in the baseline.

To guarantee the longitudinal nature, we progressively merged all CADU instances, starting with 2007 and 2008. We used a *full outer merge* to ensure that all data belonging to the same individual, in all instances, are accurately aggregated. We considered scenarios where an individual exists in both instances or in only one. We additionally checked the *LISTOF_original_NIS* across instances, merging them into a new column in the 2007–2008 temporary database (first scenario) or keeping the existing list (second scenario).

We also address temporal changes of *family_code* and *renewal_date* as it matters to epidemiological studies and occurs regardless of the biannual re-listing process. To register changes in *family_code*, we created additional columns named *family_code_YEAR* across each year within the observed period. If an individual exists in both instances (2007 and 2008, for example), we move the existing values from the corresponding instances to *family_code_2007* and *family_code_2008* and replace the *family_code* attribute by these new columns in the baseline. If an individual exists in only one instance, we populate the proper *family_code_YEAR* and keep the other empty. The same applies to *renewal_date*.

The original baseline has 15 attributes (n) prior to merge. Each merge introduces two additional columns (c) for *family_code* and *renewal_date*. So, the resulting baseline for i instances will have approximately $i*c+n$ columns. Following multiple discussions with clinicians and epidemiologists, a total of 92 fields were identified to form the cohort profile (baseline + data to be analyzed). The current cohort size (2007–2015) has 114 million records.

C. Record linkage pipeline

The linkage between CADU and PBF is deterministic, based on *current_NIS* and *LISTOF_original_NIS*, and retrieves all payments received by each cohort participant, storing these "exposure data" within the cohort profile. As there are no common key attributes among the health databases, we use a probabilistic 4-stage pipeline comprising a) data quality assessment, b) data conditioning, c) record linkage, and d) accuracy assessment.

Data quality is responsible for analyzing the input databases and identifying attributes suitable for linkage, considering their coexistence in other databases, the percentage of missing values, and their ability to uniquely identify individuals. We use the following attributes: *name, date_ofBirth, gender, mother_name* and *municipality_code*.

Data conditioning encompasses three tasks: i) data cleaning and standardization, ii) blocking and iii) data anonymization using Bloom filters. We performed data transformation and cleaning over the selected linkage attributes through the standardization of dates and names, as well the definition of default values for missing values.

Blocking is used to group records with equal values for a given attribute into blocks and perform comparisons only among such blocks, thus minimizing the computational effort. However, it can also potentially reduce accuracy due to typos or missing values that can prevent the insertion of a given record into the right block. To improve effectiveness, we split the attributes *name* and *date_ofBirth* and use a "predicate" (set of attributes): *(first_name ^ municipality_code) v (surname ^ year_ofBirth)*.

Data anonymization is based on Bloom filter [18], which is a binary vector of size n initialized with 0 (zero). The filter is composed by attributes, each one with a "weight" that corresponds to the amount of bits it occupies in the filter. Attributes are decomposed in "bigrams" (pairs of characters,

including spaces) and each bigram passes through hash functions that determine the position in the filter that must be changed from 0 to 1. Bloom filters are very accurate as two identical sets will always generate the same bit vector (no false positives). After evaluating different vector sizes and weights, we defined a 110-bit filter built from two hash functions and the attributes and weights: *name* (50 bits), *date_ofBirth* (40 bits) and *municipality_code* (20 bits).

We implemented a two-step **record linkage** process composed by a *full probabilistic* method (Figure 1), based on similarity index, and an *hybrid approach* (Figure 2), based on a mixture of deterministic and probabilistic rules. The full probabilistic method is based on the Sørensen-Dice index, given by $Dice = (2 * h) \ / \ (a \ + \ b)$, where h is the number of 1's at the same positions in both filters, and a and b the number of 1's in the first and the second filters, respectively. When compared, a *Dice=1* means filters completely equal, decreasing to 0 (zero) if there are differences. In this work, we normalized the index between 0 and 10.000.

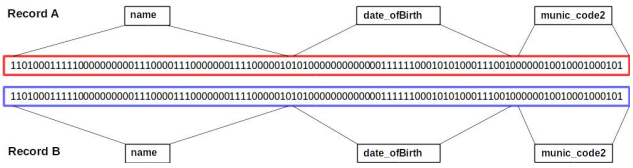


Figure 1: Full probabilistic approach.

The absence of gold standards is a key challenge to **accuracy assessment**. Since we are unable to predict the number of individuals co-existing across databases, we need to choose some cut-off points when using Dice. We experimented different cut-off points and evaluated their accuracy, observing the following: with $Dice > 8.700$, we obtained a significant number of matched pairs (true positives), but also some possibly-matched pairs (false positives). When increased to 9.200, the amount of false positives is barely any. So, we use these values (8.700 and 9.200) as lower and upper cut-off points, respectively. We manually reviewed all records encapsulated between these cut-off points to assert the effectiveness of our results.

We implemented a hybrid method based on deterministic and probabilistic rules to better improve accuracy. We use deterministic comparisons between categorical attributes or those with finite values, such as *gender* and *municipality_code*. Names and dates are probabilistically compared, as they are more sensitive to errors, and classified as: exact ($Dice=10.000$), strong ($10.000 > Dice >= 9.000$), weak ($9.000 > Dice >= 8.000$), and unpaired ($8.000 > Dice$).

These values are similar to those used in the full probabilistic approach. The difference is our hybrid approach performs individual comparisons between identical attributes and uses a decision tree to make an informed decision based on a set of predefined rules. For example, for records with

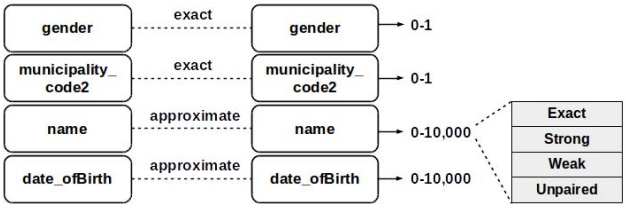


Figure 2: Hybrid approach.

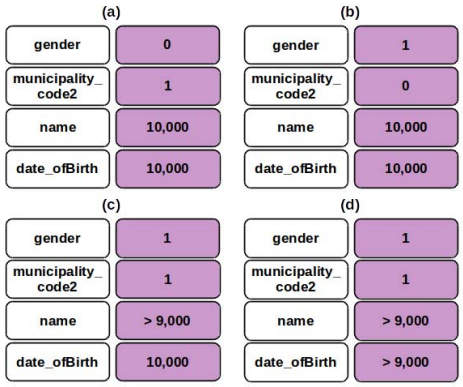


Figure 3: Example rules for the hybrid approach.

a different gender value all other attributes must match, whereas for records with identical gender values, inconsistencies between other attributes are allowed since the majority are "exact" or "strong" (see Figure 3).

Previous versions of our methods were tested with controlled databases (known coexistence of a given record) and incremental samples [19], [20], providing very accurate results. We selected incremental samples to perform linkage and assess accuracy based on sensitivity, specificity and positive predictive value (PPV) [3]. We also do manual reviews depending on sample sizes. This approach does not generate a gold standard, but enables us to validate our methods by considering the chosen cut-off points and used them in uncontrolled scenarios composed by larger samples from databases with unknown relationships.

IV. Current Results

Our linkage experiments were executed with controlled databases and manual review of records tagged as false positives. Then, we increased to larger samples from the CADU cohort and health databases.

A. Controlled scenario

We use a database with 486 records of children treated for diarrhea with positive tests for rotavirus, added to 200 other records randomly taken from other database. The second database had 9.678 records of children treated for diverse diseases, including diarrhea. The idea was to check the correct retrieval of all the 486 records among the 9.678 ones.

To replicate a real context, we used four simulation scenarios (*Si*) with different proportions (%) of character changes (letters and positions) in the attributes *name* and *date_ofBirth*. We evaluate both routines (full and hybrid) with and without blocking. Table II shows the amount of matched pairs retrieved in each scenario.

Table II: Accuracy — rotavirus.

	S1 (10,3%)	**S2 (11,3%)**	**S3 (10,3%)**	**S4 (5,15%)**
Full (no blocking)	482	481	479	482
Full (blocking)	444	332	466	458
Hybrid (no blocking)	482	482	480	486
Hybrid (blocking)	482	482	472	486

We observe that blocking tends to reduce accuracy, specially for the full probabilistic routine. Such influence is smaller in the hybrid approach, as we use predicates for blocking and perform individual comparisons of similar attributes. When we consider only no blocking results, we see that the full probabilistic routine is also quite accurate.

This experiment also provides sensitivity and PPV values to support the choice of suitable cut-off points. Table III shows the values obtained in scenario *S1*. With *Dice=8.600*, the sensitivity is 91.4% with blocking and 99.0% without blocking, with a PPV of 100%. The next Dice (8.800) has close values, suggesting that a cut-off point between 8.600 and 8.800 can be used for both metrics. The other scenarios were also analyzed to compare cut-off points and define which values we should use to all. Based on our results, we have chosen 8.700 and 9.200 as lower (true negatives) and upper (true positives) cut-off points, respectively, being all records between these values classified as false positives and subject to manual review, depending on sample sizes.

Table III: Sensitivity and PPV (full probabilistic, *S1*).

	Blocking		**No blocking**	
Dice	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**
10,000	69.3	100.0	8.8	100.0
9,800	71.2	100.0	12.8	100.0
9,600	75.3	100.0	59.5	100.0
9,400	79.4	100.0	86.6	100.0
9,200	82.3	100.0	95.3	100.0
9,000	86.4	100.0	98.1	100.0
8,800	91.4	100.0	98.8	100.0
8,600	91.4	100.0	99.0	100.0
8,400	91.4	100.0	99.2	99.8
8,200	91.4	100.0	99.2	99.8
8,000	91.4	100.0	99.2	99.8
7,000	91.4	100.0	99.2	98.2

B. Uncontrolled scenario

After using controlled scenarios, we tested our methods with cohort samples to observe their scalability and accuracy. We performed a year by year (2007 to 2011) analysis linking cohort records to mortality records (SIM database) from three Brazilian states (SE, SC and RO) with variable data quality and number of individuals in CADU. We tested other databases (hospitalizations and notifiable diseases) and calculate sensitivity and PPV for each case.

Fig. 4 shows the overall cut-off points providing better results to each sample. The maximum value below the curve (a) has reached 01 with accuracy up to 100%. The minimum value was 9.99 (c), with 97% of accuracy. We have also compared our methods without and with a second round of comparison, which nominated as "AtyImo v1" and "AtyImo v2". It is possible to observe the significant improvement that we obtain when a second round is used.

V. Conclusions and future work

We have dedicated three years designing our methods and evaluating their accuracy in controlled and uncontrolled scenarios. We observed the need of using different cut-off points even for different samples from the same database, and that manual review of dubious records is really impracticable in large volumes of data. These issues complicate the definition of gold standards for probabilistic linkage, especially in our 114 million context. In parallel, we have addressed several key challenges to build a huge population-based cohort and ensure its suitability for the desired studies.

Currently, we are working on machine learning techniques to improve accuracy and try to eliminate manual review. We are also porting our linkage methods to CUDA-capable hardware in order to use highly scalable parallel architectures without the need of blocking, which we believe can improve accuracy and reduce execution time. From the epidemiological standpoint, we started to extract data marts and apply some statistical approaches to analyze them.

Acknowledgment

This work is supported by CNPq, FINEP, FAPESB, Brazilian Ministry of Health, Bahia State Government, Wellcome Trust, Bill & Melinda Gates Foundation, Royal Society UK, Newton Fund, and UK Medical Research Council.

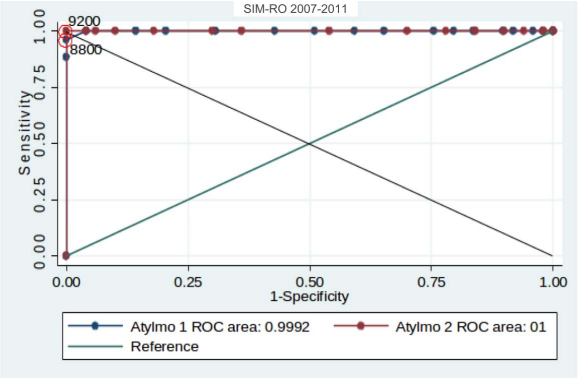
References

[1] A. Doan, A. Halevy, and Z. Ives, *Principles of data integration*, 1st ed. Morgan Kaufmann, 2012.

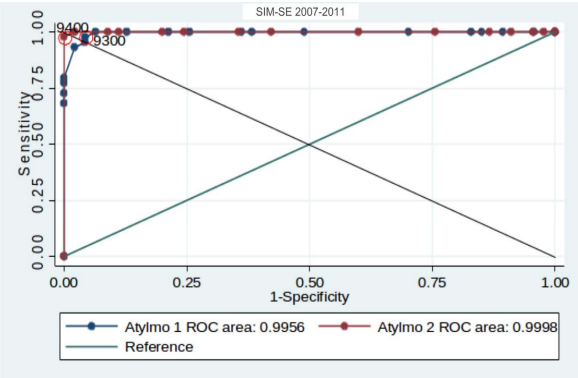
[2] I. Carneiro, *Introduction to Epidemiology: understanding Public Health*, 2nd ed. Open University Press, 2011.

[3] N. Davis and B. Shiland, *Statistics and data analytics for health data management*, 1st ed. Elsevier, 2016.

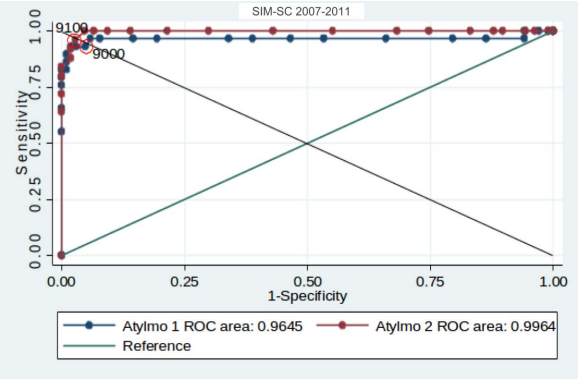
[4] A. Reeves, S. Basu, M. McKee, D. Stuckler, A. Sandgren, and J. Semenza, "Social protection and tuberculosis control in 21 European countries." *Lancet Infectious Diseases*, vol. 14, no. 11, pp. 1105–1112, 2014.



(a) Accuracy for state 1 (Rondônia (RO)).



(b) Accuracy for state 2 (Sergipe (SE)).



(c) Accuracy for state 3 (Santa Catarina (SC)).

Figure 4: Accuracy of CADU cohort X SIM.

[5] M. Pujades-Rodriguez *et al.*, "Heterogeneous associations between smoking and a wide range of initial presentations of cardiovascular disease in 1937360 people in England: lifetime risks and implications for risk prediction," *International Journal of Epidemiology*, vol. 44, no. 1, p. 129, 2015.

[6] M. F. Lima-Costa, L. C. Rodrigues, M. L. Barreto, M. Gouveia, and B. L. Horta, "Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (Epigen Initiative)." *Nature Scientific Reports*, vol. 5, no. 9812, pp. 1–7, 2015.

[7] D. Rasella, M. Harhay, R. Pamponet, Marina Aquino, and M. Barreto, "Impact of primary health care on mortality from heart and cerebrovascular diseases in brazil: a nationwide analysis of longitudinal data," *BMJ*, vol. 349, 2014.

[8] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, 1969.

[9] P. Christen and K. Goiser, "Quality and complexity measures for data linkage and deduplication," *Quality Measures in Data Mining*, vol. 43, pp. 127–151, 2007.

[10] P. Christen, *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*, 1st ed. Springer, 2012.

[11] S. M. Randall, A. M. Ferrante, and J. Semmens, "The effect of data cleaning on record linkage quality," *BMC Medical Informatics and Decision Making*, vol. 13, 06 2013.

[12] R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," *BMC Medical Informatics and Decision Making*, vol. 9, no. 41, 2009.

[13] E. Durham, Y. Xue, M. Kantarcioglu, and B. Malin, "Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage," *Information Fusion*, vol. 13, no. 4, pp. 245–259, 2012.

[14] G. Hagger-Johnson, K. Harron, A. Gonzalez-Izquierdo, M. Cortina-Borja, N. Dattani, B. Muller-Pebody, R. Parslow, R. Gilbert, and H. Goldstein, "Identifying possible false matches in anonymized hospital administrative data without patient identifiers," *Health Services Research*, 2014.

[15] K. R. d. Camargo Jr. and C. M. Coeli, "Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage," *Cadernos de Saúde Pública*, vol. 16, pp. 439 – 447, 06 2000.

[16] P. Jurczyk, J. J. Lu, L. Xiong, J. D. Cragan, and A. Correa, "Fril: A tool for comparative record linkage," *AMIA Annual Symposium Proceedings*, vol. 2008, pp. 440–444, 2008.

[17] P. Christen, "Febrl: An open source data cleaning, deduplication and record linkage system with a graphical user interface," in *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, ser. ACM SIGKDD. New York, NY, USA: ACM, 2008, pp. 1065–1068.

[18] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, pp. 422–426, 1970.

[19] R. Pita, C. Pinto, P. Melo, M. Silva, M. Barreto, and D. Rasella, "A Spark-based workflow for probabilistic record linkage of healthcare data." *EDBT/ICDT Workshops*, pp. 17–26, 2015.

[20] C. Pinto, R. Pita, P. Melo, S. Sena, and M. Barreto, "Correlação probabilística de bancos de dados governamentais," in *Simpósio Brasileiro de Bancos de Dados (SBBD 2015)*, ser. SBBD 2015. Porto Alegre, Brazil: SBC, 2015, pp. 77–85. [Online]. Available: http://dexl.lncc.br/sbbd2015/anais/Proceedings.pdf