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}@gmail.com
, 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 now considered essential to capture and model longitudinal and social aspects of diseases. However, such integration has significant challenges associated with it. Databases are often 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 associated metadata. Novel computational methods are required to integrate such databases and enable their statistical analyses for clinical 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 other administrative databases from various government entities in order to facilitate epidemiological research. We discuss and evaluate the design and validation of our data integration model and probabilistic data linkage methods for creating research data marts that can be statistically analyzed.**

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

I. Introduction

Data integration is an crucial component across several diverse application domains (research, finance, government etc) as it enables the capture and analysis of large volumes of heterogeneous data. 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) [**?**]. The former is generally 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 ongoing 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 explore and analyze the data. The primary aim of the project is the creation and validation of a very large population-based cohort, comprised by 114 million individuals (>50% of Brazil population) 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.

Due to the lack of a common, persistent 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 standard data, the use of probabilistic linkage approaches mandates the design and evaluation of specialized metrics in order to assess the accuracy of results.

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 our methods for data quality assessment and data harmonization (transformation, cleansing, anonymization and blocking). Finally, we present some experiments and discuss our performance and accuracy results.

This paper is organizing as follows: Section **??** presents some related work on record linkage tools and cohort-based initiatives. Section **??** 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 **??**, emphasizing accuracy and scalability. Finally, we present some conclusions and future works in Section **??**.

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 to find data patterns, such as done in the ALSPAC[1] and

[1]http://www.bristol.ac.uk/alspac

CONCORD[2] projects, as well in [**?**] and [**?**]. Regarding the Brazilian databases we use, there are diverse cohort-based and ecological studies, such as [**?**], [**?**], [**?**], [**?**], [**?**], 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 of record linkage, based on the comparison of attributes present in records from different data sources, was proposed by [**?**]. It is also widely discussed in [**?**] and grounded further development [**?**].

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 [**?**] is a good reference for data preparation issues. In [**?**], the authors conclude that data cleaning can represent up to 75% of the linkage effort.

Several proposals related to privacy preservation using Bloom filters exist, such as [**?**], [**?**], and [**?**]. Blocking and indexing methods are discussed in [**?**].

There are several tools for probabilistic record linkage currently available, proposed both by the academy and the industry. RecLink [**?**] was a pioneer proposal targeted to Brazilian databases. German RLC[3], Frill [**?**] and Febrl [**?**] 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.

III. Proposed approach

The overarching aim of this work was to develop and evaluate novel deterministic and 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*" (domain-specific data) for several epidemiological studies to be conducted, after a rigorous accuracy assessment step.

[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. There are two versions of CADU: version 6 (from 2007 to 2010), in which data are organized in two table groups (residences and individuals) with 167 attributes; and version 7 (2011 onwards), with 18 tables (additional data on income, work, homeless and disable people, family changes etc) totalizing 433 attributes.

Bolsa Família is the most well known welfare programme. Individuals registered in CADU and considered poor (according to specific criteria) are eligible to 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 maintains two main strategies to provide free access to health services. Despite being managed by the same department, data from these strategies are stored in approximately 40 disparate databases (ranging from administrative health data to disease-specific registers, as well organ donation, tissues banks, and transplant registers), all of which have different format, structure and variable data quality.

Common problems associated with these databases include: a) high-rates of missing data for specific groups such as homeless people or young children; b) inconsistent coding and recording patterns; c) the absence of a single, unified, unique and persistent participant identifier that spans health and administrative datasets. These challenges have significant implications with regards to the selection of fields to be utilized in the probabilistic integration of these sources.

B. Research cohort setup

The cohort must comprise all individuals registered in CADU, between 2007 and 2015, whose received at least one payment (PBF) within this period. To build it, we need to address 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 with different number of tables and attributes, as summarized in Table **??**. We use only data from residence and individuals to build the cohort. In version 6, table *A* (residence) has 42 and table *B* (individuals) has 107 attributes, respectively. In version 7, table *1* (residence) also has 42 attributes, while table *4* (individuals) has 38.

Table I: Dimensions of CADU tables

Year	Table	File Size (GB)	Number of records	Version
2007	A	11,4 GB	21.028.364	V6
	B	86.8 GB	79.050.446	
2008	A	12.5 GB	22.767.472	
	B	100.1GB	89.915.568	
2009	A	13.5 GB	24.661.693	
	B	108.8 GB	97.640.845	
2010	A	14.3 GB	26.107.223	
	B	114.4 GB	102.663.287	
2011	1	25 GB	27.014.194	V7
	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	

Common attributes to both versions were iteratively evaluated and included in a *inner merge* based on *family_code*. Data normalization routines (e.g. data conversion, adjustment of categorical variables) were applied. As a 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, original_NIS, registration_date, registration_status, municipality_code, renewal_date*.

As registration in CADU must be renewed biannually, we need to deal with multiple NIS assigned to the same individual that occurs due to several reasons. Individuals change their family, due to marriage or divorce, receiving a new *family_code* that keeps assigned to their NIS. For registration purposes, NIS 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 $n = 15$ attributes prior to merge. Each merge introduces two additional columns ($c = 2$) 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 is 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. Regarding the health databases, as there are no common key attributes, we must use a probabilistic approach based on a 4-stage pipeline covering a) data quality assessment, b) data conditioning, c) record linkage, and d) accuracy assessment.

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

Data conditioning encompasses three main components: 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 build the "predicate" (set of attributes): *(first_name ˆ municipality_code) v (surname ˆ year_ofBirth)*.

Data anonymization is based on Bloom filter [**?**], which is a binary vector of size n initialized with 0 (zero). The filter is composed by a set of 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 sets composed by the same attributes will always generate the same bit vector (no false positives). After iteratively evaluating different vector sizes (n) and weights, we defined a 110-bit filter built from two hash functions and the following attributes (and weights): *name* (50 bits), *date_ofBirth* (40 bits) and *municipality_code* (20 bits).

2) Pairwise comparison: We implemented a two-step data linkage process composed of 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 the filters are 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.

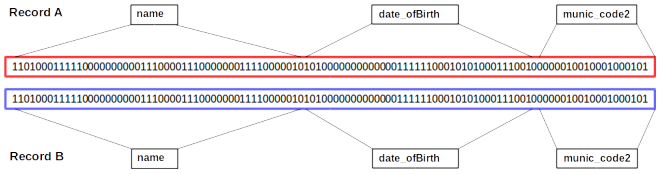


Figure 1: Full probabilistic approach.

The absence of gold standards is a key challenge to assess accuracy. In our case, 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 to decide if two records match. We performed experiments with different cut-off points and assessed the accuracy, observing the the following situation: with a $Dice > 8.700$, 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 in order to assess the accuracy and present the results.

To improve accuracy, we implemented a hybrid method based on deterministic and probabilistic rules. 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 ($8.000 > Dice$).

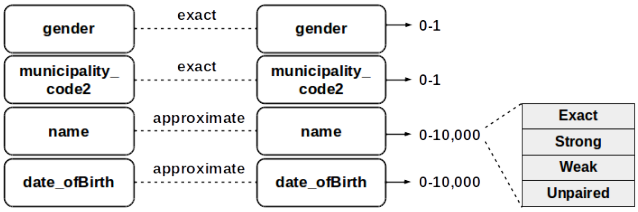


Figure 2: Hybrid approach.

These values are similar to those used in the full probabilistic approach. The difference is that the 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 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". Figure 3 (b)–(d) depicts some possible combinations.

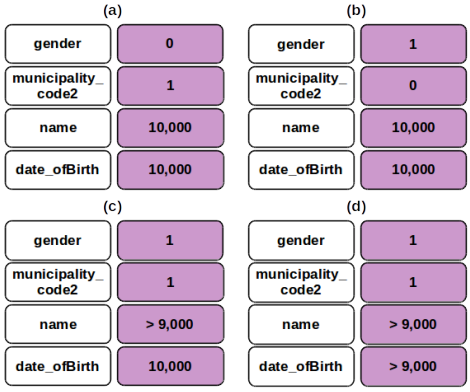


Figure 3: Example rules for the hybrid approach.

The full probabilistic routine is more flexible, as we can adjust the cut-off points in order to get more or less matched records. This routine retrieves a larger number of matched records, however with a significant number of false positives. The hybrid approach is more restrictive: it brings a smaller set of matched records with a high number of true positives.

3) Accuracy assessment: We must consider two key problems: the absence of gold standards (no cut-off points suitable for all cases) and the very large amount of data that makes manual review of records impractical.

Previous versions of our methods were tested with controlled databases and incremental samples [**?**], [**?**], providing

very accurate results. Controlled databases are databases for which we can infer the coexistence of a given record. We selected incremental samples to perform the linkage and assess the accuracy based on sensitivity, specificity and positive predictive value (PPV) [**?**]. We also do manual reviews depending on the 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 while the cohort setup process was taking place. We performed tests with controlled databases and manual review of records tagged as false positives to assess the accuracy of our probabilistic routines. Then, we increased to larger samples from the 2011 instance of CADU and health databases.

A. Controlled scenario

We use a database with 486 records of children treated in hospitals for diarrhea with positive tests for rotavirus, added to 200 other records randomly taken from other database. The second database has 9.678 records of children treated for diverse diseases, including diarrhea, in these hospitals. The idea was to evaluate if our routines correctly retrieve 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 **??** shows the amount of matched pairs (true positives) retrieved in each situation.

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 step also aims to aid in the choice of suitable cut-off points to our linkage routines, based on sensitivity and PPV. To exemplify, Table **??** shows the values of sensitivity and PPV obtained with the full probabilistic routine in scenario *S1*. With *Dice=8.600*, we have a sensitivity of 91.4% with blocking and 99.0% without blocking, with a PPV of 100%. The next higher value (8.800) has very 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 scenarios. Based on our results, we have chosen 8.700 and 9.200 as lower and upper cut-off points, respectively. All records with a Dice between these values are considered false positives and subject to manual review, depending on the sample size. Above 9.200, the records are classified as true positives.

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

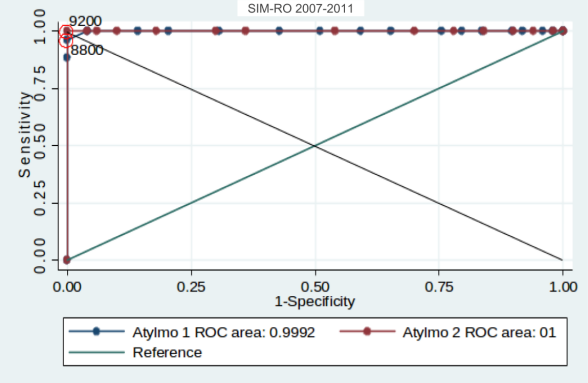
After these experiments in controlled scenarios, we tested our methods with samples from the generated cohort to observe their scalability and accuracy. We performed a year by year (2007 to 2011) analysis linking cohort records to mortality (SIM database) records from three different Brazilian states (SE, SC and RO), with variable data quality and number of individuals registered in CADU. We performed tests with 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.

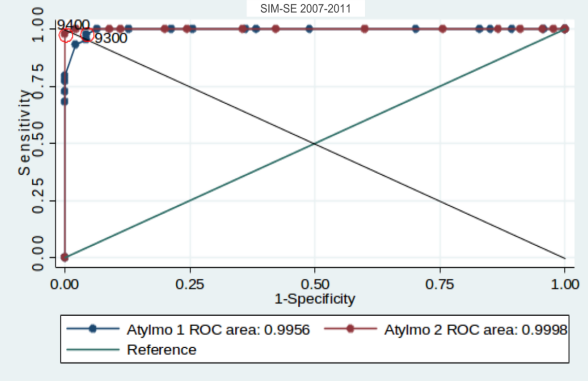
Our current implementation is based on Spark and runs over computational clusters scaling up to 64 nodes, with linkage times up to 8 hours, depending on the databases involved. We are also evaluating a parallel implementation of our methods targeted to hybrid parallel architectures comprised by multi-GPU hardware, which is able to link up to 20 million records in around 25 seconds.

V. Conclusions and future work

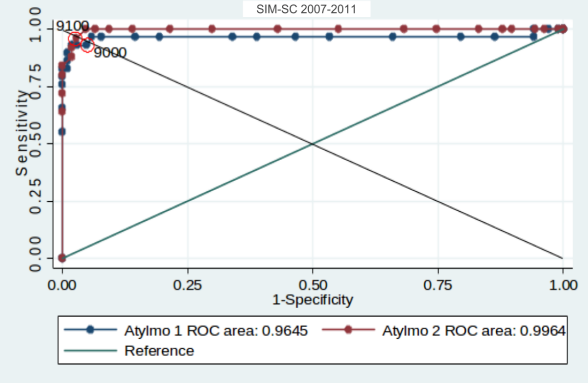
We have dedicated almost three years designing our methods and evaluating their accuracy in controlled and



(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 for (a) SE, (b) SC and (c) RO states using AtyImo with one and two rounds.

uncontrolled scenarios. 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. 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-based hardware in order to use highly scalable parallel architectures without the need of using blocking, which we believe can improve accuracy and reduce the execution time. From the epidemiological standpoint, we started to extract data marts and apply some statistical approaches to analyze these data.

Acknowledgment

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

References

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

[2] A. Reeves, S. Basu, M. McKee, D. Stuckler, A. Sandgren, and J. Semenza, "Social protection and tuberculosis control in 21 european countries, 1995-2012: a cross-national statistical modelling analysis." *Lancet Infectious Diseases*, vol. 14, no. 11, pp. 1105–1112, 2014.

[3] M. Pujades-Rodriguez, J. George, A. D. Shah, E. Rapsomaniki, S. Denaxas, R. West, L. Smeeth, A. Timmis, and H. Hemingway, "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.

[4] M. F. Lima-Costa, L. C. Rodrigues, M. L. Barreto, M. Gouveia, B. L. Horta, J. Mambrini, F. S. G. Kehdy, A. Pereira, F. Rodrigues-Soares, C. G. Victora, and E. Tarazona-Santos, "Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling brazilians (the epigen initiative)." *Nature Scientific Reports*, vol. 5, no. 9812, pp. 1–7, 2015.

[5] B. L. Horta, D. Gigante, H. Gonçalves, J. V. dos Santos Motta, C. L. de Mola, I. Oliveira, F. Barros, and C. GVictora., "Cohort profile update: The 1982 pelotas (brazil) birth cohort study." *International Journal of Epidemiology*, vol. 44, no. 2, pp. 441–441e, 2015.

[6] J. Gaspar, T. Sá, Z. Reis, R. Júnior, M. Júnior, and R. G. ao, "Use of geographic information system tools in research on neonatal outcomes in a maternity-school in belo horizonte - brazil," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 5, no. 12, pp. 91–96, 2014.

[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] J. S. Nery, S. M. Pereira, D. Rasella, M. L. F. Penna, R. Aquino, L. C. Rodrigues, M. L. Barreto, and G. O. Penna, "Effect of the brazilian conditional cash transfer and primary health care programs on the new case detection rate of leprosy," *PLOS Neglected Tropical Diseases*, vol. 8, no. 11, pp. 1–7, 11 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pntd.0003357

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

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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. [Online]. Available: http://dx.doi.org/10.1016/j.inffus.2011.04.004

[16] 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*, pp. n/a–n/a, 2014. [Online]. Available: http://dx.doi.org/10.1111/1475-6773.12272

[17] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE Trans. on Knowl. and Data Eng.*, vol. 24, no. 9, pp. 1537–1555, Sep. 2012.

[18] 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.

[19] 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.

[20] P. Christen, "Febrl -: An open source data cleaning, deduplication and record linkage system with a graphical user interface," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 1065–1068. [Online]. Available: http://doi.acm.org/10.1145/1401890.1402020

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

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

[23] 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

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