On the Accuracy and Scalability of Probabilistic Data Linkage over the Brazilian 114 Million Cohort

Robespierre Pita, Clícia Pinto, Samila Sena, Rosemeire Fiaccone,
Leila Amorim, Sandra Reis, Mauricio L. Barreto, Spiros Denaxas, Marcos E. Barreto

***Abstract*—Data linkage refers to the process of identifying and linking records that refer to the same entity across multiple heterogeneous data sources. This method has been widely utilized across scientific domains, including public health where records from clinical, administrative and other surveillance databases are aggregated and used for research, decision-making, and assessment of public policies. When a common set of unique identifiers do not exist across sources, probabilistic linkage approaches are used to link records using a combination of attributes. These methods require a careful choice of comparison attributes as well as similarity metrics and cut-off values to decide if a given pair of records matches or not and for assessing the accuracy of the results. In large, complex datasets, linking and assessing accuracy can be challenging due to the volume and complexity of the data, the absence of a gold standard and the challenges associated with manually reviewing a very large number of record matches. In this paper, we present AtyImo, a hybrid probabilistic linkage tool optimized for high-accuracy and scalability in massive datasets. We describe the implementation details around anonymization, blocking, deterministic and probabilistic linkage and accuracy assessment. We present results from linking a large population-based cohort of 114 million individuals in Brazil to public health and administrative databases for research. In controlled and real scenarios, we observed high accuracy of results: 93%–97% true matches. In terms of scalability, we present AtyImo's ability to link the entire cohort in less than nine days using Spark and scaling up to 20 million records in less than 12 seconds over heterogeneous (CPU+GPU) architectures.**

***Keywords*-Data linkage, Accuracy assessment, Cohort study.**

I. Introduction

DATA linkage is a widely adopted technique for combining data from disparate heterogeneous sources potentially belonging to the same entity [1], [2]. It has been applied in several domains to aggregate data to be used in decision-making processes, monitoring and surveillance tasks, assessment of public policies and clinical research [3], [4].

In the context of public health, we linked data from a very large socioeconomic cohort consisting of 114 million individuals who have received payments from a conditional cash transfer programme in Brazil between 2007 and 2015 to records from public health databases. We generated bespoke data sets for research studies aiming to quantify and evaluate the impact of such payments on several disease outcomes.

Besides data volume, the complexity of our scenario comes from the absence of common key attributes in all databases involved. This imposes the use of probabilistic approaches which, in turn, have a strong requirement on accuracy. Another challenging issue is the lack of gold standards to validate these linkages, as the amount of cohort participants appearing in any health database is unknown.

This is an extended version of our award-winning poster presented at IEEE Biomedical and Health Informatics 2017 [5]. In this paper, we present our data linkage tool (AtyImo) and its pipeline structure for anonymization, block construction and pairwise comparison. AtyImo implements a mixture of deterministic and probabilistic routines for data linkage. We discuss and evaluate accuracy, scalability and performance results achieved in experimental and real scenarios.

This paper is organized as follows: Section II presents some related work on data linkage tools and accuracy assessment issues. Section III presents the AtyImo tool and describes its functionalities. We provide a summary of our case study in Section IV. Different accuracy and scalability results are presented and discussed in Section V, and some conclusions and ideas for further research are presented in Section VI.

II. Related work

Data linkage is implemented in vendor-specific databases and analytics platforms, statistical software and research-centered solutions. In this section, we list some existing tools relying on some form of block construction and probabilistic technique to enable data linkage. Additionally, we discuss the accuracy assessment process and its associated challenges.

A. Data linkage tools

Reclink [6] provides different matching and block construction routines to support data linkage. Phonetic codes are applied over linkage attributes to generate candidate blocks for pairwise comparison. It was used in some ecological and small-size (nearly 5,700 individuals) cohort-based studies using Brazilian governmental data, such as [7] and [8].

Merge Toolbox (MTB) [9] is offered by the German Record Linkage Center together with other tools for privacy-preserving matching (Safelink) and error imputation for accuracy validation (TDGen). FRIL [10] offers an interactive

R. Pita, C. Pinto and M. E. Barreto are with the Institute of Mathematics and Statistics, Computer Science Department, Federal University of Bahia, Salvador, Brazil, e-mail: (pierre.pita, cliciasp, marcosb)@ufba.br.

S. Sena, R. Fiaccone and L. Amorim are with the Institute of Mathematics and Statistics, Department of Statistics, Federal University of Bahia, Salvador, Brazil, e-mail: (mylasenna, fiaccone, leiladen)@ufba.br.

S. Reis and M. L. Barreto are with the Centre for Data and Knowledge Integration for Health (CIDACS), Oswaldo Cruz Foundation, Salvador, Brazil, e-mail: (ssreis, mauricio.barreto)@bahia.fiocruz.br.

S. Denaxas and M. E. Barreto are with the Institute of Health Informatics, University College London, London, UK, e-mail: (s.denaxas, m.barreto)@ucl.ac.uk. M. E. Barreto is a Royal Society's Newton International Fellow (2016–2018).

Manuscript received May 15, 2017.

linkage process allowing users to select comparison attributes, a similarity function and a decision model to accept or reject matched records. Febrl [11] is an open-source tool with a graphical interface that allows the combination of different encoding, indexing, comparison and classification functions.

HARRA [12] and NC-Link [13] are proposals focused on machine learning techniques to perform record classification of large-scale data sets. Machine learning-driven approaches are also used in [14] to classify clusters of records generated by the MFIBlocks algorithm for uncertain multi-entity resolution, as well in [15] for classifying online customer profiling data. Different meta-blocking algorithms to entity resolution are discussed in [16], with emphasis on load balancing, graph (block) construction and entity comparison.

A data mining platform targeting health care data is presented in [17]. It employs Apache Drill to support schemaless access to diverse data sources. Authors claim that this platform shortens the time needed to make data available for analysis when compared to other existing tools, presenting runtime performance results for *join* and *distinct* queries.

In [18], parallel data linkage algorithms and performance results obtained with data sets scaled up to 6 million records are discussed. Further, in [19], a Web-based version of these algorithms is compared against Febrl and FRIL. Privacy-preserving linkage methods implemented in OpenCL are discussed in [20], with emphasis on block construction and similarity calculation. Different blocking and clustering techniques to scale record linkage methods are discussed in [21]. Hybrid architectures were used in [22] to evaluate linkage performance over NVIDIA and OpenCL, resulting on a speedup of 10 times for a 1.7 million data set from *freedb.org*.

B. Accuracy of data linkage

Common approaches for accuracy assessment comprise of: i) the usage of "gold standards" (when true match status is known), ii) sensitivity analysis based on different linkage criteria, iii) comparison between linked and non-linked records, and iv) statistical techniques dealing with uncertainty and bias measurement [23]. The entire scope of this topic also comprises proposals dealing with data quality and preparation, multiple imputation problems, bias and uncertainty quantification, as well scalability modeling [24].

When gold standards are absent, one must rely on controlled experiments with small size databases from which we can perform manual review on linked records to quantify accuracy and then scale to bigger databases. Accuracy can be measured through sensitivity, specificity, positive predictive value (PPV) and receiver operating characteristic (ROC) curves, as discussed in [25], [26], and [27].

An alternative approach to assessing the accuracy is to utilize machine learning techniques for automating the process of tuning the linkage hyperparameters and reduce or eliminate the amount of human intervention. This approach is discussed in [28] and delivers highly accurate results from unsupervised methods as compared to existing gold standards. In [29], a discussion is provided on the manner that artificial neural networks and clustering algorithms can be used to deal with missing data and produce accurate results.

Our work contributes to the field by providing a scalable tool capable of linking very large databases with complex relationships and great variability in terms of data quality. Other contributions arise from the discussion on metrics for accuracy assessment, reference cut-off values and establishment of gold standards for probabilistic linkage.

III. AtyImo data linkage tool

We initially developed AtyImo in 2013 to serve as a linkage tool supporting a joint Brazil–UK project aiming at to build a large population-based cohort with data from more than 100 million participants and produce disease-specific data to facilitate diverse epidemiological research studies. The volume and heterogeneity of the databases involved, as well the absence of common key attributes among them and the expected cohort size (initially 80 million records) have posed strong requirements on scalability and accuracy. To address these challenges, we designed and implemented AtyImo as a modular pipeline, encapsulating components for data preprocessing, pairwise comparison and matching decision.

Prior to linkage, all input data sets pass through a data quality analysis stage which performs data integrity and missingness checks, which quantify the percentage of missing data especially from linkage attributes. Any required procedures for data cleansing are also applied in this stage. The goals are to identify the suitability of linkage attributes (given missing data statistics) and recover records presenting some imputation errors that can be fixed through standardization procedures. The processes have been implemented using a variety of statistical analysis tools such as Stata and R.

A. Data preprocessing

This stage is responsible for data harmonization, blocking and anonymization. Common operations for data harmonization comprise date and string formatting, removal of special characters and insertion of specific values for missing data.

Blocking [23] is a common approach used in data linkage that consists of grouping candidate pairs with similar characteristics to be subsequently compared. It reduces the execution time, as only similar blocks (and not all existing ones) are compared at a time. Blocking additionally helps to overcome space (memory) limitations when dealing with large databases.

Different techniques can be used for blocking: single key, predicates or machine learning-driven methods. Single key blocking is simpler, as only one attribute is used to group records, but errors in this key attribute can prevent a given record to be inserted into a block, thus never being compared to potential pairs. A refined approach consists of combining several key attributes into a disjunctive predicate used to correctly block records even if some of these attributes have errors. Finally, classification algorithms can use specific rules to learn how effectively and accurately construct blocks.

Figure 1 shows the predicate-based blocking strategy used in AtyImo. After analyzing different predicates, we have chosen the following one: *(name AND mother_name AND municipality_code) OR (surname AND mother_surname AND year_of_birth)*. It guarantees that errors in one clause do not

prevent the record to be correctly grouped, resulting in a relatively small number of blocks of moderate size.

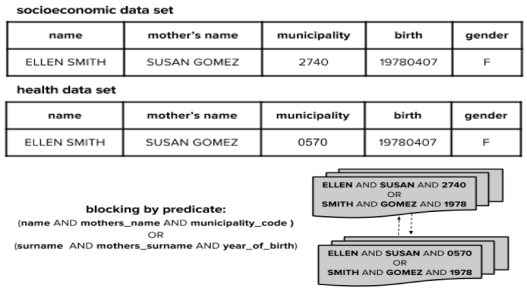

Fig. 1. Blocking predicate implemented by AtyImo.

Anonymization is a critical issue for health data and different privacy-preserving techniques can be used to address this problem [30], [31], [32]. AtyImo uses Bloom filters [33], which are binary vectors of size *n* initialized with 0 (zero). Linkage attributes being anonymized are decomposed in "bi-grams" (pairs of characters, including spaces) processed by hash functions to determine which positions in the filter must change to 1, as depicted in Figure 2. The amount of positions depends on each attribute's weight. Bloom filters are very reliable as two identical set of attributes will always generate the same vector (no false positives). After evaluating different configurations, we defined a 180-bit filter built from two hash functions and the following attributes (and weights): *name* and *mother_name* (50 bits each), *date_of_birth* (40 bits), *municipality_code* and *gender* (20 bits each).

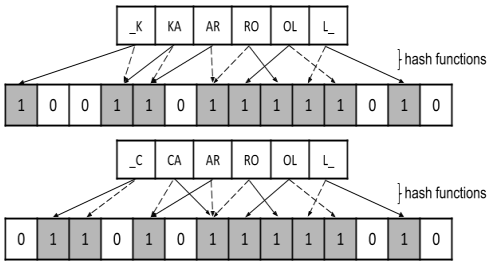

Fig. 2. Example of a Bloom filter encoding hashed bigrams.

B. Pairwise comparison methods

AtyImo provides two approaches for pairwise comparison. The first one is *full probabilistic*, in which Bloom filters representing linkage attributes are entirely compared (Figure 3). A similarity value is calculated based on the Sørensen-Dice index [34], defined as $Dice = (2 * h) \ / \ (a \ + \ b)$, being *h* the total of 1's at the same positions in both filters, and *a* and *b* the total of 1's in the first and second filters, respectively. A *Dice=1* means filters completely equal, decreasing to 0 (zero) depending on existing differences. Our implementation normalizes Dice indices between 0 and 10,000.

The second approach is a *hybrid* mixture of deterministic and probabilistic rules applied to individual linkage attributes (Figure 4). Categorical attributes are matched exactly, whereas

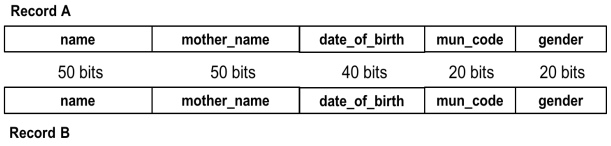

Fig. 3. Full probabilistic linkage approach comparing Bloom filters directly.

names and dates (both more prone and sensitive to errors) are probabilistically classified as: exact (Dice=10,000), strong (10,000 > Dice >= 9,000), weak (9,000 > Dice >= 8,000), and unpaired (8,000 > Dice). This approach results in some flexibility on the combinations of exact and approximate comparisons.

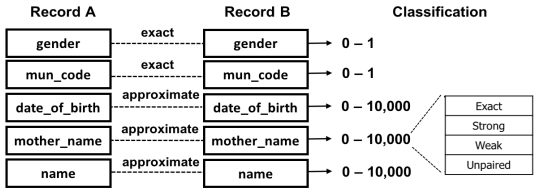

Fig. 4. Hybrid linkage approach based on bespoke rules.

These methods produce three output data sets: true positive pairs, true negative pairs and "dubious records" (false positive and false negative matches). This classification is based on upper and lower cut-off points representing boundaries for true positive and true negative matches, respectively. We perform an analysis on which cut-off points retrieve more true (positive and negative) pairs and perform an iterative second round over dubious pairs, shifting these points in each iteration, to retrieve additional records into these two groups.

C. Accuracy assessment

Resulting data sets produced by AtyImo are evaluated based on sensitivity, specificity and positive predictive value. We perform manual review over samples with incremental size from controlled databases (known coexistence of matching records) used as gold-standards. This process allows us to quantify AtyImo's accuracy, especially with regards to the choice of cut-off points that minimize the amount of false pairs. Additonally, it enables our approach to scale up to bigger databases (as in our case study), where manual review is impossible or impractical.

Current Dice indices used as upper and lower cut-off points are 9,400 and 8,800, respectively. We chose the cut-off points after several iterative tests with different samples of variable size and data quality extracted from different databases. These tests aimed to check the variation on indices providing better results and the possibility of using the same values for all linkage executions. Table I summarizes one of our results obtained with four cohort samples linked to hospital episodes (SIH) and disease notification (SINAN) databases. We observed Dice values providing better accuracy varying between 9,100 and 9,400, which highlights the challenges of trying to establish a default value.

TABLE I
VARIABILITY OF BEST DICE COEFFICIENTS.

	SIH			SINAN		
Samples	**Dice**	**Sens.**	**PPV**	**Dice**	**Sens.**	**PPV**
SE	9,400	95.6%	95.0%	9,300	96.7%	95.9%
SC	9,100	99.0%	96.0%	9,100	97.7%	97.4%
BA	9,100	98.5%	97.9%	9,200	95.7%	95.5%
RO	9,300	94.1%	94.2%	9,400	87.9%	91.0%

We have started testing machine learning methods to design an automated accuracy checker and automatically retrieve dubious records. As we need to scale up to 114 million records, we expect this approach to help us in eliminating the need for manual review and to efficiently deal with the variability of Dice values. Some preliminary results are discussed in [35].

IV. CASE STUDY: THE 114 MILLION COHORT

The Brazilian 114 million cohort [36] is a joint Brazil-UK effort started in 2013 with the aim of building a population-based cohort to enable diverse research studies on disease epidemiology and surveillance. The cohort was constructed based on data from CadastroÚnico (CADU database), a central register for individuals intending to participate in more than 20 social and protection programmes kept by the Brazilian government. Bolsa Família (PBF database) is one of these programmes and provides conditional cash transfers to families considered poor or extremely poor. So far, the cohort is comprised of 114 million individuals who have received payments from Bolsa Família between 2007 and 2015. This cohort is linked to public health databases to generate disease-specific data used in epidemiological studies.

Linkages between CADU and databases from social programmes (including PBF) are deterministic, based on the NIS number, a unique identifier similar to a social security number. Linkages between the cohort and public health databases (the main ones are summarized in Table II) are performed probabilistically, as there are no common key identifiers across these databases. We developed AtyImo to enable us to perform these linkages in an accurate fashion.

TABLE II
GOVERNMENTAL DATABASES.

Databases	**Coverage**
CADU (socioeconomic data)	2007 to 2015
PBF (cash benefits payments)	2007 to 2015
SIH (hospitalizations)	1998 to 2011
SIM (mortality)	2000 to 2012
SINAN (notifiable diseases)	2000 to 2010
SINASC (live births)	2001 to 2012

V. ACCURACY AND SCALABILITY RESULTS

Our evaluation strategy to assess AtyImo's accuracy and scalability was based on some small size, controlled databases (where the number of matching pairs was known), as well samples from the CADU cohort with increasing sizes and variable data quality. We calculated accuracy metrics for each case and used ROC curves to visualize which cut-off points are the best similarity threshold discriminating matching pairs. Depending on sample sizes, we additionally performed manual review for checking the results obtained and their accuracy.

AtyImo is implemented over Spark and over heterogeneous (CPU+multi-GPU) architectures. Synthetic data sets and both implementations are publicly available[1]. The Spark-based implementation is structured as nine Python modules summarized in Table III. The *correlation()* module is the most time-consuming as it performs pairwise comparisons, similarity calculations and matching decisions.

TABLE III
ATYIMO-SPARK CODE ORGANIZATION.

Module	**Purpose**
preprocessing.py	Data cleansing and standardization
createBlockKey.py and *writeBlocks.py*	Blocking (record grouping)
encondingBlocking.py	Creation of Bloom filters
correlation.py	Pairwise comparison and matching
dedupByKey.py and *createDatamart.py*	Generation of research datasets
config.py and *config_static.py*	Data and Spark configuration

A. Accuracy in controlled scenarios

Table IV presents a comparative analysis linking a controlled database with positive tests for rotavirus (children treated for diarrhoea) to a database with children's hospital admissions for all-cause diseases (including diarrhoea). The first database had 486 records, to which we added 200 additional random records as noise. The second database had 9,678 records. The goals were to correctly retrieve all 486 records from the second database (simulating a controlled behaviour) and compare AtyImo's results against other tools.

TABLE IV
COMPARATIVE ANALYSIS – ATYIMO X FRIL X FEBRL.

	FRIL	**FRIL blocking**	**Febrl**	**Febrl blocking**	**AtyImo**	**AtyImo blocking**
TP	486	484	480	479	486	486
TN	0	0	0	0	0	0
FP	1	0	1	0	0	0
FN	0	2	6	7	0	0

We observed similar accuracy in terms of true positive (TP) and true negative (TN) pairs, with a slight advantage for AtyImo when considering false positive (FP) and false negative (FN) pairs. We used the same comparison strategy for FRIL and Febrl: attributes *name* and *mother_name* were compared through the Jaro-Winkler distance (weight = 1), date difference for *date_of_birth* (weight = 0.9), exact match for *municipality_code* and *gender* (weight = 0.8 for both). This configuration is similar to AtyImo's hybrid approach. Blocking was based on the *sorted neighborhood algorithm*, which sorts records through a given key and only compares records within a predefined distance window, whereas for AtyImo we used the predicate described in Section III. As FRIL and Febrl have a black-box implementation, we were unable to fully explore how blocking influences the results obtained.

[1] https://github.com/spiros/atyimo

B. Accuracy in uncontrolled scenarios

While the cohort creation was taking place, we performed experiments linking isolated CADU samples (from 2007 to 2015) to health databases covering specific diseases (e.g. tuberculosis, children mortality, BCG vaccination etc). Table V presents linkage results for tuberculosis between the CADU 2011 (best quality sample), the hospitalizations (SIH) and the disease notifications (SINAN) databases. We used samples from two Brazilian states: Sergipe (SE), the smallest sample (few individuals in CADU), and Santa Catarina (SC), a middle size sample. They were chosen for manual review purposes.

TABLE V
LINKAGE RESULTS (SAMPLE: CADU TUBERCULOSIS 2011).

Databases (number of records)	**Matched pairs**		**True positives (%)**	
	Full	Hybrid	Full	Hybrid
CADU 2011 x SIH SE (1,447,512) x (49)	40	24	23 (57.5%)	23 (95.8%)
CADU 2011 x SIH SC (1,988,599) x (330)	140	95	83 (59.2%)	86 (90.5%)
CADU 2011 x SINAN SE (1,447,512) x (624)	398	311	309 (77.6%)	299 (96.1%)
CADU 2011 X SINAN SC (1,988,599) x (2,049)	661	500	551 (83.3%)	462 (92.4%)

The hybrid approach retrieved more true positive pairs compared to the full probabilistic routine, which emphasizes individual comparison of linkage atributes provides more accurate results less influenced by imputation erros. We made similar tests with a bigger sample (BA) and a poorest data quality sample (RO) (Table I).

In [5], we presented the overall cut-off points providing better results when linking cohort records to different mortality (SIM) samples (RO, SE and SC), respectively: 9,300 (sensitivity 94.3%, PPV 95.9%), 9,300 (sensitivity 97.6%, PPV 97.7%), 9,000 (sensitivity 86.6%, PPV 93.5%). We plotted ROC curves for all experiments to visually assess the power of discrimination of each coefficient, as depicted in Figures 5 to 7. Results from the SC sample were slightly worse compared to other samples, having been influenced by expressive missing data present in 2007 to 2009 fragments.

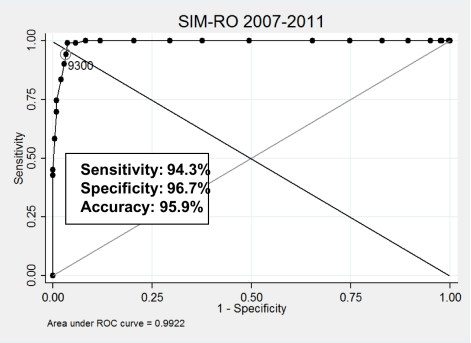

Fig. 5. Best coefficient and related results (CADU cohort x SIM, RO).

From these experiments, we observed which Dice values provided the best results for each case and measured the distance between them to verify the suitability of using the same coefficients for all linkages. Best coefficients varied from 8,800 to 9,400, being used as thresholds to separate dubious records. This observed variation reinforces the complexity of running probabilistic linkages without gold standards.

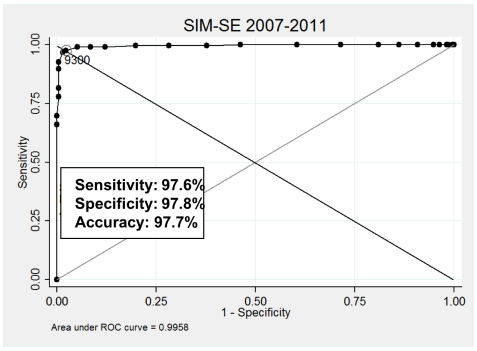

Fig. 6. Best coefficient and related results (CADU cohort x SIM, SE).

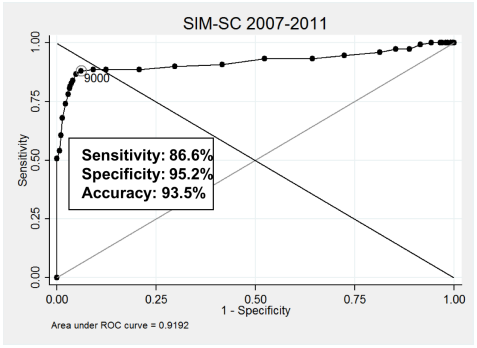

Fig. 7. Best coefficient and related results (CADU cohort x SIM, SC).

C. Scalability evaluation

We measured the time spent on linkage for each tool in Table IV. Average times (in seconds) for five executions were: FRIL (681), Febrl (3780), AtyImo (103); decreasing to FRIL (37), Febrl (2730), AtyImo (42) using blocking. Although these results were obtained with a small database, they illustrate how AtyImo performs as good as other tools. We consider AtyImo's major advantage as its ability to scale upwards to huge databases, which we were unable to do with other tools. We linked the entire cohort to 370,000 records from SINAN in nine days using 20 nodes (40 2.8GHz cores, 256GB RAM) from a dedicated supercomputer. We also linked 7 million cohort records to one million records from SIM in four days using a 56-core (3.1GHz, 512GB RAM) server.

Considering the potential speed up of parallel architectures, we have ported AtyImo to heterogeneous (CPU+GPU) platforms aiming to simultaneously use all available processors to distribute data and tasks. We have used a static strategy to assign data and tasks over available CPU and GPU subsystems. We initialize the runtime with as many CPU threads as CUDA devices, since one CPU thread is linked to each GPU to perform memory and control operations, plus a number of CPU threads linked to each CPU core to perform multicore computation. Each group of processing elements executing a

computational kernel is seen as a combined processing unit, since CPU and GPU threads work in a coordinated fashion.

Scalability tests were performed for the *correlation* function since it is the most time-consuming component within the pipeline. Files to be linked are loaded in two matrices with one line per record. We exploit parallel matrix calculation and perform summation by partitioning the outermost loop into independent, variable size chunks, which allow us to better distribute the workload. Algorithm 1 shows the parallel version of AtyImo, where *cpu_exec()* and *kernel()* correspond to CPU and GPU versions, respectively, of the *correlation.py* module.

Algorithm 1 AtyImo code using OpenMP and CUDA.

```
INPUT
matrixA       // larger matrix (dataset A)
matrixB       // smaller matrix (dataset B)
matrizA_gpu   // matrixA chunk at GPU
matrizA_cpu   // matrixA chunk at CPU
nlines_a      // # of lines of matrixA
nlines_b      // # of lines of matrixB
num_col       // # of columns in both matrices
puThreshold   // matrixA chunk in each processor
qtd_gpu       // available GPUs

OUTPUT: Dice (similarity) between records

1: int *puThreshold = getPuThreshold(qtd_gpu,
                      percentage_each_gpu);
2: omp_set_nested(1);
3: omp_set_num_threads(num_gpus);
4: #pragma omp parallel num_threads(qtd_gpu+1)
5: {
6:     int id = omp_get_thread_num();
7:     if(id == 0) {
8:        int *matrixA_cpu = split(matrixA,
                              puThreshold);
9:       #pragma omp parallel
          num_threads(threads_cpu) {
10:          int idNested = omp_get_thread_num();
12:          cpu_exec(matrixA_cpu, matrixB,
             nlines_b, puThreshold, idNested);
13:       }
14:   }
15:    else if(id != 0) {
16:       cudaSetDevice(id);
17:       cudaGetDevice(&gpu_id);
18:       int *matrixA_gpu;
19:       matrizA_gpu = split(matrixA,
                        puThreshold);
20:       kernel(matrixA_gpu, matrixB,
          nlines_a_gpu, nlines_b, num_col);
21:   }
22:}
```

Figure 8 illustrates the execution time and speed up obtained with samples varying from 1 to 20 million records linked using one or two GPUs, as well hybrid CPU+GPU cores. Speed up was calculated based on the CPU cores subsystem. The maximum speed up was around 8 for the hybrid subsystem, with a sustained ability to scale up to 20 million records. Our platform comprised of 4 Intel Xeon processors (3.33GHz, 100GB RAM, 6 cores and 128MB cache each) and 2 Tesla C2070 GPUs (448 cores in total). We used CUDA version 7.5.

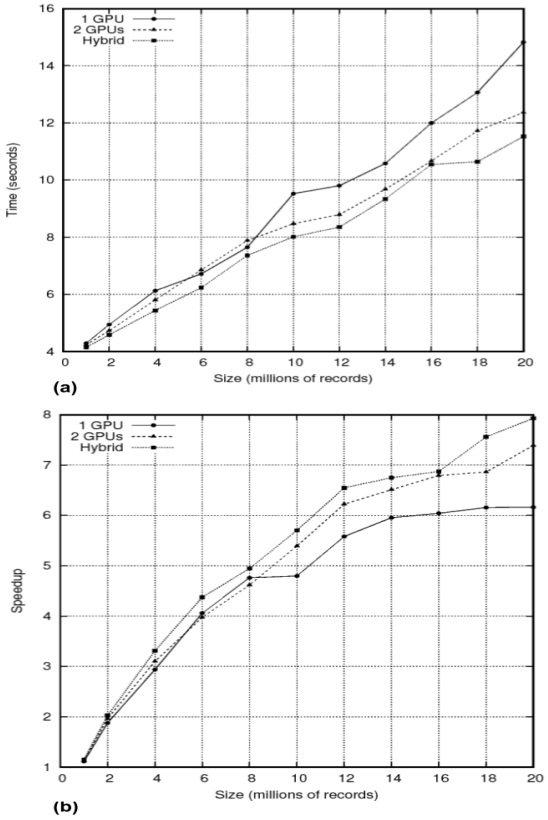

Fig. 8. Execution time (a) and speed up (b) – AtyImo hybrid.

VI. Conclusion

In this paper, we described and evaluated our probabilistic linkage approach implemented by AtyImo through accuracy and scalability results achieved in controlled and uncontrolled experiments. We analyzed accuracy metrics and ROC curves to identify effective similarity indices to generate high-accurate data for epidemiological studies. The variability of best Dice indices emphasized the complexity regarding the definition of gold standards for probabilistic linkage. AtyImo has proved to be very accurate linking controlled and uncontrolled databases. Its major contribution is the ability to link huge databases within a reasonable execution time and with good accuracy.

We are working on improvements that will enable AtyImo to operate for the entire cohort over GPU architectures and designing machine learning methods to automatic accuracy assessment. We consider a careful discussion on the quality of data linkage as essential, particularly when large databases are used and mismatches might have important consequences for statistical analyzes in terms of bias.

Acknowledgments

This study was funded by CNPq, FINEP, FAPESB, Bill & Melinda Gates Foundation (OPP1161996) and The Royal Society (NF160879). It was also supported by the National Institute for Health Research (RP-PG-040710314), Wellcome Trust (086091/Z/08/Z), and the Farr Institute of Health Informatics Research, funded by The Medical Research Council (MR/K006584/1), in partnership with Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the

Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government) and the Chief Scientist Office (Scottish Government Health Directorates).

REFERENCES

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

[2] E. Herrett, A. D. Shah, R. Boggon, S. Denaxas, L. Smeeth, T. van Staa, A. Timmis, and H. Hemingway, "Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study," *BMJ*, vol. 346, p. f2350, 2013.

[3] S. C. Denaxas, J. George, E. Herrett, A. D. Shah, D. Kalra, A. D. Hingorani, M. Kivimaki, A. D. Timmis, L. Smeeth, and H. Hemingway, "Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER)," *International Journal of Epidemiology*, vol. 41, no. 6, pp. 1625–1638, 2012.

[4] S. C. Denaxas and K. I. Morley, "Big biomedical data and cardiovascular disease research: opportunities and challenges," *European Heart Journal–Quality of Care and Clinical Outcomes*, vol. 1, no. 1, pp. 9–16, 2015.

[5] C. Pinto, R. Dantas, S. Sena, S. Reis, R. Fiaccone, L. Amorim, M. Barreto, S. Denaxas, and M. Barreto, "Accuracy of probabilistic linkage: the Brazilian 100 million cohort," in *International Conference on BIomedical and Health Informatics*, ser. BHI 2017. Orlando, USA: IEEE/EMBS, 2017, pp. xx–xx.

[6] K. R. d. Camargo Jr. and C. M. Coeli, "Reclink: an application for database linkage implementing the probabilistic record linkage method," *Cadernos de Saúde Pública*, vol. 16, pp. 439 – 447, 06 2000.

[7] 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.

[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: https://doi.org/10.1371/journal.pntd.0003357

[9] 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.

[10] 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.

[11] 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.

[12] H.-s. Kim and D. Lee, "Harra: Fast iterative hashed record linkage for large-scale data collections," in *Proceedings of the 13th International Conference on Extending Database Technology*, ser. EDBT 2010. New York, NY, USA: ACM, 2010, pp. 525–536. [Online]. Available: http://doi.acm.org/10.1145/1739041.1739104

[13] Y. Jeon, J. Yoo, J. Lee, and S. Yoon, "Nc-link: A new linkage method for efficient hierarchical clustering of large-scale data," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2017.

[14] T. Sagi, A. Gal, O. Barkol, R. Bergman, and A. Avram, "Multi-source uncertain entity resolution at yad vashem: Transforming holocaust victim reports into people," in *Proceedings of the 2016 International Conference on Management of Data*, ser. SIGMOD '16. New York, NY, USA: ACM, 2016, pp. 807–819. [Online]. Available: http://doi.acm.org/10.1145/2882903.2903737

[15] C. Conrad, N. Ali, V. Keelj, and Q. Gao, "Elm: An extended logic matching method on record linkage analysis of disparate databases for profiling data mining," in *2016 IEEE 18th Conference on Business Informatics (CBI)*, vol. 01, Aug 2016.

[16] V. Efthymiou, G. Papadakis, G. Papastefanatos, K. Stefanidis, and T. Palpanas, "Parallel meta-blocking for scaling entity resolution over big heterogeneous data," *Information Systems*, vol. 65, no. Supplement C, pp. 137 – 157, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030643791530199X

[17] E. Begoli, T. Dunning, and C. Frasure, "Real-time discovery services over large, heterogeneous and complex healthcare datasets using schema-less, column-oriented methods," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, March 2016, pp. 257–264.

[18] A.-A. Mamun, T. Mi, R. Aseltine, and S. Rajasekaran, "Efficient sequential and parallel algorithms for record linkage," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 252–262, 2014.

[19] A.-A. Mamun, R. Aseltine, and S. Rajasekaran, "RLT-S: A web system for record linkage," *PLOS ONE*, vol. 10, no. 5, pp. 1–9, may 2015.

[20] Z. Sehili, L. Kolb, C. Borgs, R. Schnell, and E. Rahm, "Privacy preserving record linkage with ppjoin," in *Datenbanksysteme für Business, Technologie und Web (BTW)*, 2015, pp. 85–104.

[21] S. Rendle and L. Schmidt-Thieme, *Scaling Record Linkage to Non-uniform Distributed Class Sizes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 308–319.

[22] B. Forchhammer, T. Papenbrock, T. Stening, S. Viehmeier, and U. D. F. Naumann, "Duplicate detection on GPUs," in *BTW*. Köllen-Verlag, 2013, pp. 165–184.

[23] K. Harron, H. Goldstein, and C. Dibben, *Methodological developments in data linkage*, 1st ed. John Wiley & Sons, 2016.

[24] H. Goldstein, K. Harron, and M. Cortina-Borja, "A scaling approach to record linkage," *Statistics in Medicine*, vol. 36, no. 16, pp. 2514–2521, 2017, sIM-16-0569.R3. [Online]. Available: http://dx.doi.org/10.1002/sim.7287

[25] P. Christen and K. Goiser, *Quality and Complexity Measures for Data Linkage and Deduplication*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 127–151. [Online]. Available: https://doi.org/10.1007/978-3-540-44918-8_6

[26] M. A. Bohensky, D. Jolley, V. Sundararajan, S. Evans, D. V. Pilcher, I. Scott, and C. A. Brand, "Data linkage: A powerful research tool with potential problems," *BMC Health Services Research*, vol. 10, no. 1, p. 346, Dec 2010. [Online]. Available: https://doi.org/10.1186/1472-6963-10-346

[27] R. Aldridge, K. Shaji, A. Hayward, and I. Abubakar, "Accuracy of probabilistic linkage using the enhanced matching system for public health and epidemiological studies," *PLOS ONE*, vol. 10, no. 8, pp. 1–15, 08 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0136179

[28] S. J. Grannis, J. M. Overhage, S. Hui, and C. J. McDonald, "Analysis of a probabilistic record linkage technique without human review," *AMIA Annual Symposium Proceedings*, vol. 2003, pp. 259–263, 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479910/

[29] G. P. Hettiarachchi, N. N. Hettiarachchi, and D. S. Hettiarachchi, "Next generation data classification and linkage: Role of probabilistic models and artificial intelligence," in *IEEE Global Humanitarian Technology Conference (GHTC 2014)*, Oct 2014, pp. 569–576.

[30] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *SIGKDD Explor. Newsl.*, vol. 10, no. 2, pp. 12–22, Dec. 2008. [Online]. Available: http://doi.acm.org/10.1145/1540276.1540279

[31] T. Gál, G. Kovács, and Z. Kardkovács, "Survey on privacy preserving data mining techniques in health care databases," *Acta Universitatis Sapientiae, Informatica*, vol. 6, no. 1, pp. 33–55, 2014.

[32] S. Sathya and T. Sethukarasi, "Efficient privacy preservation technique for healthcare records using big data," in *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, Feb 2016, pp. 1–6.

[33] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Commun. ACM*, vol. 13, no. 7, pp. 422–426, Jul. 1970.

[34] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[35] R. Pita, E. Mendonça, S. Reis, M. Barreto, and S. Denaxas, "A machine learning trainable model to assess the accuracy of probabilistic record linkage," in *Big Data Analytics and Knowledge Discovery*, L. Bellatreche and S. Chakravarthy, Eds. Cham: Springer International Publishing, 2017, pp. 214–227.

[36] D. Rasella, R. Aquino, C. A. Santos, R. Paes-Sousa, and M. L. Barreto, "Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of brazilian municipalities," *The Lancet*, vol. 382, no. 9886, pp. 57–64, 2017/12/12.