International Journal of Medical Informatics 114 (2018) 45–51

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

A comparison of accuracy and computational feasibility of two record linkage algorithms in retrieving vital status information from HIV/AIDS patients registered in Brazilian public databases

Adelzon Assis de Paula[a,*], Denise Franqueira Pires[b], Pedro Alves Filho[b], Kátia Regina Valente de Lemos[b], Eduardo Barçante[c], Antonio Guilherme Pacheco[a]

[a] PROCC/FIOCRUZ, Avenida Brasil, 4365, Rio de Janeiro, Brazil
[b] Rio de Janeiro State Health Secretariat, Rua México, 128, Rio de Janeiro, Brazil
[c] DataUERJ/UERJ, Rua São Francisco Xavier, 524, Rio de Janeiro, Brazil

ARTICLE INFO

Keywords:
Record linkage
Deterministic linkage
Probabilistic linkage
HIV
Public datasets
Mortality

ABSTRACT

Background and objective: While cross-referencing information from people living with HIV/AIDS (PLWHA) to the official mortality database is a critical step in monitoring the HIV/AIDS epidemic in Brazil, the accuracy of the linkage routine may compromise the validity of the final database, yielding to biased epidemiological estimates. We compared the accuracy and the total runtime of two linkage algorithms applied to retrieve vital status information from PLWHA in Brazilian public databases.

Methods: Nominally identified records from PLWHA were obtained from three distinct government databases. Linkage routines included an algorithm in Python language (PLA) and Reclink software (RlS), a probabilistic software largely utilized in Brazil. Records from PLWHA[1] known to be alive were added to those from patients reported as deceased. Data were then searched into the mortality system. Scenarios where 5% and 50% of patients actually dead were simulated, considering both complete cases and 20% missing maternal names.

Results: When complete information was available both algorithms had comparable accuracies. In the scenario of 20% missing maternal names, PLA[2] and RlS[3] had sensitivities of 94.5% and 94.6% ($p > 0.5$), respectively; after manual reviewing, PLA sensitivity increased to 98.4% (96.6–100.0) exceeding that for RlS ($p < 0.01$). PLA had higher positive predictive value in 5% death proportion. Manual reviewing was intrinsically required by RlS in up to 14% register for people actually dead, whereas the corresponding proportion ranged from 1.5% to 2% for PLA. The lack of manual inspection did not alter PLA sensitivity when complete information was available. When incomplete data was available PLA sensitivity increased from 94.5% to 98.4%, thus exceeding that presented by RlS (94.6%, $p < 0.05$). RlS spanned considerably less processing time compared to PLA.

Conclusion: Both linkage algorithms presented interchangeable accuracies in retrieving vital status data from PLWHA. RlS had a considerably lesser runtime but intrinsically required manually reviewing a fastidious proportion of the matched registries. On the other hand, PLA spent quite more runtime but spared manual reviewing at no expense of accuracy.

1. Introduction

Record linkage refers to the process of matching information from different datasets corresponding to the same individual or entity [1]. Record linkage involves two critical steps: a) a searching routine in which potentially linkable information are brought together for inspection and b) a comparison to infer whether the information referred to on each record are in fact from the same unit [2].

Though a number of routines are currently available to deal with more challenging scenarios [3], three main linkage types are broadly recognized. Manual linkage is the strategy by which records from two separate sources are manually compared and deemed as true matches

* Corresponding author at: PROCC/FIOCRUZ, Residência Oficial, Avenida Brasil, 4365, Manguinhos, CEP: 21040-360, Brazil.
E-mail addresses: adelzon@hotmail.com (A.A. de Paula), denise.pires@saude.rj.gov.br (D.F. Pires), pedro.filho@saude.rj.gov.br (P.A. Filho), katia.valente@saude.rj.gov.br (K.R.V. de Lemos), eduardo.barcante@uerj.br (E. Barçante), apacheco@fiocruz.br (A.G. Pacheco).

[1] People living with HIV/AIDS

[2] Python Language Algorithm

[3] Reclink software

https://doi.org/10.1016/j.ijmedinf.2018.03.005
Received 9 August 2017; Received in revised form 19 March 2018; Accepted 19 March 2018

or not [4]. Manually linking records might suffice for some simplistic applications but it becomes time consuming and even unpractical as the amount of data become larger [5].

Deterministic linkage routines are based on exact-match comparisons of either one univocal identifier common to both databases or a combination of variables (e.g. name, surname and date of birth) to yield unique discrimination [6]. Deterministic routines range from simply linking datasets by a univocal identifier to more refined stepwise approaches allowing variation between pairing records [3]. Probabilistic routines, on the other hand, rely on weighting matches and non-matches based on error probabilities and frequency distributions of attributed values in the input databases [5], taking into account the degree to which two matches comply with the agreement and disagreement weights for each identifier [7].

The choice of a suitable algorithm or its combined application depends on aspects such as the proportion of erroneous entries and missing values, the actual discriminating power of the identifiers and the prior knowledge on the database's completeness [7]. As a rule of thumb, whenever good quality direct identifiers are available deterministic algorithms are preferable; conversely, probabilistic routines are indicated when such identifiers are not available or when data is of poor quality [7]. In practical terms, however, this decision is left to the users and is based on their preferences and their ultimate goals for the linkage project [3].

Currently, as both the diversity of information sources and the length of individual datasets increase, the efficiency of record linkage algorithms is considered to be better approached not only on the basis of accuracy measures but also considering its computational feasibility, thus accounting for the time elapsed while processing [8].

Irrespective of the routine applied, record linkage is being increasingly utilized to enable health researchers to gather longitudinal information for entire populations [5,9]. Information provided by health care delivery system and by monitoring and surveillance, constitutes a major source of data on both mortality and morbitidy [10], which can be further integrated into a larger comprehensive database for epidemiological and research purposes.

In Brazil, the linkage of various public databases is instrumental in monitoring the HIV/AIDS epidemic [11]. Notwithstanding the diversity of sources of information, merging data to the official mortality database (Mortality Information System/SIM) constituted our primary interests as routine searches in the SIM are performed to both identify deaths among those HIV/AIDS cases mandatorily reported and to detect unreported cases [12].

While linking information from PLWHA to the official mortality databases is a critical step in monitoring the HIV/AIDS epidemic in the country [11], the accuracy of the linkage routine may compromise the validity and generalizability of the final merged database, yielding to biased estimates [3]. False-positive matches, that is, erroneously matching records that pertain to distinct individuals, can both underestimate survival and overestimate disease incidence among external cohorts linked to the mortality registry [13,14]; false-negative non-matches, by contrast, can bias risk differences and the risk ratios toward the null value [13].

From the variety of linkage algorithms currently available to assist retrieving vital status information from PLWHA two are of primary interest in the present analyses, because they have been used to link databases of HIV/AIDS patients with other public databases in a regular basis. RlS, a probabilistic approach-based software routinely used by the national AIDS program to link public databases [15,16] and a previously validated deterministic algorithm, used on a regular basis to retrieve vital status from patients lost to follow-up in a HIV/AIDS cohort [17–22].

Though both routines have been extensively assessed in terms of diagnostic accuracy, to our knowledge no comparative assessment of their accuracies and computational feasibilities has been carried out so far. Therefore, it would be of great value to critically examine the potentialities of such algorithms in cross-referencing information from PLWHA to the mortality database so as to determine the most suitable application strategy, in terms of single or combined utilization and runtime processing aiming to improve HIV/AIDS case surveillance and to assist researchers to accurately gather information from public databases.

In the present manuscript, we compared the accuracy and the total runtime of two linkage algorithms in linking information from PLWHA registered in HIV/AIDS public databases to the SIM database.

2. Materials and methods

2.1. Data sources and inclusion criteria

We employed data from three distinct sources: the Medication Logistics Control System (SICLOM), which provides logistic support regarding antiretroviral therapy dispensation [23], the Laboratory Test Control System (SISCEL), which monitors information on laboratory tests [24] and the SIM database.

Fake test datasets containing different proportions of records from people known to be alive and from people actually dead (PAD) were assembled in order to determine sensitivity, specificity and predictive values. Data from people known to be alive consisted of information from PLWHA on antiretroviral treatment by the end of December 2012 according to SICLOM. Data from PAD comprised information from PLWHA registered as deceased in the SICLOM between January 2008 and December 2009.

Whenever information on the vital status from SICLOM and SIM diverged, data were validated using SISCEL information. Records from patients having dubious or conflicting information on the vital status were excluded.

Four different scenarios were considered; firstly, we simulated a scenario wherein PAD occurred in a 50% proportion. To this end, we assembled datasets with 200 records from PAD randomly selected combined to 200 records from PLWHA known to be alive. Alternatively, we considered a scenario of 5% PAD, consisting of a random sample of 200 records from PAD combined to 3800 records from PLWHA known to be alive. Power and sample size calculations have been described elsewhere [25]. Those test datasets were then searched in the mortality database in a time frame from January 2008 to December 2010, the outcome being defined as “finding a record in the mortality database given it is truly there.”

The two last scenarios consisted of datasets with 50% and 5% PAD associated to the randomly removal of 20% of maternal names from the test databases. Distinct datasets were constructed for each scenario so as to warrant further validity. To minimize bias, two independent researchers performed the manual review process. Total runtime was assessed for every linkage procedure as the time elapsed between session initiation and obtaining the final read-to-use output.

Importantly, information on patients' death among PLWHA on antiretroviral therapy is regularly entered into SICLOM database through a specific form (available at: http://www.aids.gov.br/pt-br/pub/2017/formulario-de-cadastramento-de-obito-siclom), thus consisting in an independent source of death information apart from SIM.

As a complementary analysis, we used information from Information System for Notifiable Diseases (SINAN). Data from PAD consisted of AIDS cases reported to SINAN through the death criterion, which is adopted when AIDS is diagnosed after the patient's death. We utilized data form SINAN between January 2008 and December 2009. Data from people known to be alive consisted of information from individuals who died in 2012 according to SIM. Again, those datasets were searched in the SIM database from January 2008 to December 2010.

2.2. Record linkage algorithms and data preprocessing

The first linkage approach constituted a deterministic algorithm written in Python programming language (PLA) [18]. PLA was primarily implemented to assist in the retrieval of information on the vital status of PLWHA lost to follow-up in two large urban HIV/AIDS cohorts [17–19] and it has been routinely used for such purpose [21,22]. PLA has also been adapted to cross-reference PLWHA public databases to both tuberculosis [19,20] and AIDS cohorts databases [21].

PLA correlates records using exact comparisons and also allows for minor errors in names and dates of birth, measured by means of phonetic codes and a string similarity score implemented with the *difflib* module (https://pymotw.com/2/difflib/), which helps dealing with specific differences between sequences and dates. Here, PLA ran both in a fully automated procedure (PLA-FAP) or associated to manual review of unresolved not excluded pairs (PLA-MR). Here we used Python 2.7.13 and *difflib* 2.3 versions. Patient's name, mother's name and date of birth were used as matching fields with parameter estimates obtained with the Expectation–Maximization algorithm. The field's name and mother's name were compared using the Levenshtein distance string comparator measure [14]

The second algorithm, Reclink software (RlS), is an open-source, multiplatform package based on probabilistic record linkage strategy built in C^{++} programming language [26]. The system was developed in the late 1990's and has been updated on a regular basis. We utilized RlS version 3.1.6.3160.

A number of manuscripts making use of RlS can be found in the literature [27–30]. A multiple-step blocking strategy was employed to maximize the finding of true matches [31], with the last two steps requiring manual review. Blocking keys applied were the phonetic codes for the patient's first and last name and sex; comparison variables and linkage parameters used included patient's name (matching probability 98% and error probability 0.0011%; threshold 85%), mother's name (matching probability 74%, and error probability 0.0046%; threshold 85%) and date of birth (matching, probability 98%, and error probability 2.356%; threshold 65%) [11,32].

2.3. Accuracy measures and statistical analysis

Sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) were calculated, with the positive case being death. Confusion matrices were constructed by cross tabulating data regarding the true status of the patients (dead or alive) and the results of the linkage procedure (matched or not matched). Differences in accuracy were assessed in a paired study design, where both algorithms were applied to the same set of records [25].

Sensitivity and specificity were compared by exact binomial test, whereas PPV and NPV were compared by relative predictive values [25]. Those measures, along with their corresponding 95% confidence intervals were estimated by means of DTComPair package. As specificity and NPV are thought to be skewed due to the large number of potential matches identified during the blocking steps [5], we also assessed f-measure, with β set at 1.0 so as to equally weight sensitivity and PPV [3].

All record linkage procedures run on a Windows 10 machine with 2.60 GHz Core i7 processor and 8 GB RAM. Statistical analyses were performed in the R software environment version 3.3.3 (https://cran.r-project.org/).

3. Results

Table 1 shows accuracy values for PLA and RlS when complete information was available. Though sensitivity for PLA-FAP increased from 96.8% to 98.9% after manual review it did not differ from RlS (98.4%, $p > 0.05$). Specificity for PLA and RlS did not differ. Increasing proportions of people actually dead had only marginal impact over NPV and PPV values for both algorithms.

In the scenario of 20% missing maternal names, PLA-FAP and RlS had quite similar sensitivities, respectively 94.5% (91.3–97.8) and 94.6% (91.3–97.9, $p > 0.5$); after manual reviewing, PLA sensitivity increased to 98.4% (96.6–100.0) exceeding that for RlS ($p < 0.01$). Though in low magnitude, both PLA had significantly higher specificities compared to RlS [99.9% (99.8–100.0) and 99.6% (99.4–99.8), $p < 0.01$], respectively.

NPP and PPV for both algorithms were virtually interchangeable despite the proportion of deaths, but PPV in 5% PAD was higher for PLA-FAP and PLA-MR compared to RlS: 97.8% (95.6–99.9), 97.9% (95.8–99.9) and 91.6% (87.7–95.6), respectively ($p < 0.01$ for both; Table 1).

F-measure ranged from 96.1% to 98.9% for PLA, mainly depending on the proportion of PAD and to a lesser degree on the amount of information available; the procedure with no manual review only slightly impacted the metrics. As for RlS, f-measure varied from 93.1% (when 5% PAD were present and data lacked 20% maternal names) to 98.9% (Tables 1 and 2).

Manual reviewing was intrinsically required by RlS in 6% to 14% of the number of PAD, respectively when full information and incomplete data were entered; as for PLA, the corresponding proportions were 1.5%–2%. The lack of manual inspection did not imply in statistically significant changes in PLA sensitivity when complete information was available. When incomplete data was available, however, manual reviewing increased PLA sensitivity from to 94.5% to 98.4%, thus exceeding that presented by RlS (94.6%, $p < 0.05$, Table 2). No noteworthy differences in the accuracy measures were found when using information from SINAN database (See Supplementary Data).

RlS processing time spanned between 0.7 h in a 5% PAD scenario with complete data and 1.5 h when 50% of truly deceased individuals were used and 20% maternal names lacked. As for PLA, processing time ranged between 10 h (50% PAD) up to 74 h (5% PAD).

4. Discussion

To the best of our knowledge this is the first study to and a previously validated algorithm implemented in Python language (PLA) and also to assess RlS in the context of fictitious datasets combined to a known and reliable gold standard for vital status. The body of data presented points to two quite accurate linkage algorithms available to the retrieval of vital status information from PLWHA registered in public databases.

Although both algorithms performed well when full information was available, in the presence of incomplete information PLA-FAP had slightly higher sensitivity compared to RlS, which translated into a significantly greater PPV when 5% records were from people actually dead (97.8% and 91.6%, respectively; $p < 0.01$). It should be emphasized that, such scenario simulates the proportion of deaths expected to occur in open cohorts of PLWHA from developing countries, which averages approximately 5/100 person-years [33].

Two major caveats are worth mentioning. First, RlS runs intrinsically in a rule-oriented supervised process which can be rather exhaustive as the blocking steps increases [31]. PLA, on the other hand, is implemented as a pre-gauged, chained pipeline [18], so not requiring complementary interactivity nor intervention to be fully carried out. While some trade-off between sensitivity and specificity may arise when introducing operational interactivity in opposition to a fully automated process [3,5], no relevant accuracy differences were seen despite PLA unsupervised performance.

Second, the decision on whether or not to incorporate manual reviewing as a complement to electronic linkage approaches. In some situations electronic algorithms are not sufficient to unambiguously judge all matches as either true or false pairs; in these scenarios manually reviewing cases may constitute an attractive option [34]. The amount of potential matches left to be manual reviewed is given by

Table 1
Accuracy criteria (95% CI) for Python language algorithm and Reclink software. Complete information was entered.

Accuracy Criteria	PLA-FAP	PLA-MR	RlS	p-value*	
				PLA-FAP	PLA-MR
Sensitivity	96.8 (94.2–99.3)	98.9 (97.4–100.0)	98.4 (96.6–100.0)	> 0.05	> 0.05
TP/FN count	194/6	198/2	197/3		
Specificity	99.0 (97.6–100.0)	99.0 (97.6–100.0)	99.5 (98.5 100.0)	> 0.05	> 0.05
TN/FP count	198/2	198/2	199/1		
50% PAD					
PPV	99.4 (98.4–100.0)	98.9 (97.4–100.0)	99.5 (98.4–100.0)	> 0.05	> 0.05
TP/FP count	194/2	198/2	197/1		
NPV	97.1 (94.8–99.4)	99.0 (97.6–100.0)	98.5 (96.8–100.0)	> 0.05	> 0.05
TN/FN count	198/6	198/2	199/3		
F-measure	98.1	98.9	98.9		
5% PAD					
PPV	96.8 (94.3–99.3)	97.9 (95.9–99.9)	96.7 (94.4 99.3)	> 0.05	> 0.05
TP/FP count	196/10	195/16	192/28		
NPV	99.9 (99.8–100.0)	99.9 (99.8–100.0)	99.7 (99.6–100.0)	> 0.05	> 0.05
TN/FN count	3790/4	3784/5	3772/8		
F-measure	96.8	98.4	97.5		

PLA-FAP: Python linkage algorithm –fully automated procedure; PLA-MR: Python linkage algorithm – manual review; RLS: Reclink software; PAD: people actually dead; PPV: positive predictive value; NPV: negative predictive value.

* When compared to RlS.

Table 2
Accuracy criteria (95% CI) for Python language algorithm and Reclink software. Data missed 20% maternal names.

Accuracy Criteria	PLA-FAP	PLA-MR	RlS	p-value*	
				PLA-FAP	PLA-MR
Sensitivity	94.5 (91.3–97.8)	98.4 (96.6–100.0)	94.6 (91.3–97.9)	> 0.05	< 0.05
TP/FN count	189/11	197/3	190/10		
Specificity	99.9 (99.8–100.0)	99.9 (99.8–100.0)	99.6 (99.4–99.8)	< 0.01	< 0.01
TN/FP count	199/1	200/0	199/1		
50% PAD					
PPV	98.9 (97.4–100.0)	98.9 (97.4–100.0)	98.9 (97.4–100.0)	> 0.05	> 0.05
TP/FP count	189/1	197/0	190/1		
NPV	98.5 (96.8–100.0)	99.0 (97.6–100.0)	98.0 (96.1–99.9)	> 0.05	> 0.05
TN/FN count					
F-measure	96.6	98.6	96.7		
5% PAD					
PPV	97.8 (95.6–99.9)	97.9 (95.8–99.9)	91.6 (87.7–95.6)	< 0.01	< 0.01
TP/FP count	190/8	194/7	189/24		
NPV	99.7 (99.6–99.9)	99.9 (99.8–100.0)	99.7 (99.6–99.9)	> 0.05	> 0.05
TN/FN count	3792/10	3793/6	3776/11		
F-measure	96.1	98.1	93.1		

PLA-FAP: Python linkage algorithm –fully automated procedure; PLA-MR: Python linkage algorithm – manual review; RLS: Reclink software; PAD: people actually dead; PPV: positive predictive value; NPV: negative predictive value.

* When compared to RlS.

factors as the databases sizes, the difficulty in classifying all the data and the intrinsic inability of some algorithms to accurately discriminate matches from nonmatches [5].

Notwithstanding its potential to add sensitivity to linkage procedures when a fairly small number of pairs is left to be scrutinized, manually reviewing pairs may introduce subjectivity to the linkage procedure and a trade-off between sensitivity and specificity may arise when larger amount of data has to be reviewed [35]. As a matter of fact, manually reviewing data is a potential source of bias and may not actually increase accuracy [36,37].

RlS intrinsically required manual inspection of up to 14% of the total records with the interest condition. Conversely, the inclusion of manual review for PLA had no significant impact on its sensitivity when complete data were used and only marginally improved the algorithm when data lacked 20% maternal names, with sensitivity increasing from 94.5% to 96.6%. One should keep in mind, however, that regarding RlS, both the accuracy and the proportion of dubious matches left for manual inspection are clearly dependent on the linkage strategy adopted, with other strategies yielding to distinct results [26].

In a previous analyses PLA exhibited high accuracy when no information was missing, which in a practical application significantly decreased the loss to follow-up rate and hence increased death rates in cohorts of PLWHA [18]. In spite of both studies having used SIM as the reference dataset, the accuracy levels presented ratify such measurements when poorer quality data was entered as the comparison dataset (in this case, containing information from SICLOM). The accuracy of RlS presented only partially agrees with a previous report [14], when the authors described 87.6% and 99.6% respectively for sensitivity and specificity, which we ascribe to the different linkage parameters and blocking strategies adopted.

With reference to processing time, RlS spent considerably less time than PLA irrespective of the proportion of people actually dead or the data completeness. Possible explanations for this finding are: first, RlS was written in a compiled language (C++), which is known to be faster and require less memory usage than interpreted languages such as Python; second, RlS development intended to optimize the system's performance [26], which is not the case for PLA.

Another reasoning is that PLA did not use the field sex in the blocking phase and thus the resulting fewer blocking keys increased the computational time. It should be noted that PLA running time is intrinsically related to cubic time for the Ratcliff-Obershelp pattern recognition found in the *SequenceMatcher* class from *difflib* module. One should consider, therefore, trading-off an increased processing time in a fully unsupervised process as for PLA over an optimized processing time spent in an interactive supervised algorithm as RlS. Additionally, modifications are readily extensible regarding PLA (e.g. adding sex a blocking key) so as to speed up its performance.

Though canonically used for assessing the quality of data matching, accuracy may be treacherous as specificity and negative predictive value are often inflated given the large number of true nonmatches identified during the blocking phase [3]. To overcome that potential issue and also to account for unbalances in the dataframes we included the f-measure, a compound metrics representing trade-offs between sensitivity and specificity [38]. As a rule of thumb, the f-measure should ideally exceed 95% [3], which was the case for both routines examined except for RlS when poorest data were made available in the presence of a low proportion of people actually dead. As for the canonical metrics of accuracy, sparing manual review only slightly decreased the f-measure of PLA.

A relevant aspect before generalizing accuracy measures presented is the assumption that registers from PLWHA with unknown vital status (that is, in real-world situations) have similar chances of being accurately matched as those used in our analyses [39]. Since we used routine information from public databases to generate the datasets, it seems reasonable to infer that the measures calculated here will also be generalizable to such registries.

It is noteworthy that PLA allows for specific errors in data, measured by both phonetic codes and a built-in string similarity score. In situations where typing mistakes occur at a high rate, as is the case in government databases [40], manual reviewing process and matching errors would potentially increase [41]; in this context, string comparison functions as such implemented in PLA can somewhat improve the accuracy of the comparison vectors and therefore the linkage quality [36], which may explain the results presented.

The RlS linkage parameters employed seem to place more emphasis on specificity as oppose to false negative rate; such strategy is based on the notion that false positive misclassification of outcomes in survival analyses, even when non-differentially related to the exposure variables, can bias both the risk difference and the risk ratio towards the null value [14]. It should be kept in mind, however, that in follow-up studies based on registry information, false-positive and false-negative results would respectively over- or underestimate survival and that the antagonistic effects of both errors partly cancel out, with net bias depending on the relative frequency of such errors [13].

Two major strengths of our study are the use of data from an independent source of death information apart from the official mortality database to simulate known proportions of individuals with the condition of interest and the utilization of a *bona fide* gold standard to which the algorithm findings were compared. A limitation of the study would be the potentially misclassified information on the vital status in the SICLOM. In fact, an important issue that may arise when linking records to death databases to ascertain patients' vital status is to single out missed matches, given it is often unknown whether the patient is actually alive or dead [42]. By validating the information on the vital status with data from the laboratory tests database, we circumvented this limitation, thus yielding to unbiased estimates.

It should be kept in mind that the Ratclif/Obershelp method implemented with *difflib* module has a suboptimal effectiveness to compare string matching, so using a different algorithm in spell check may be needed to further improve PLA accuracy. On the other hand, the Levenshtein distance calculated in Reclink processing puts more weight into single substitution in contrast to letter transposition.

Although a study revealed similar values for a probabilistic and a deterministic algorithm, the former was considered more consistent, as the sensitivity for deterministic routines may decrease when data with different identifier characteristics (greater typographical mistakes rates, for instance) are entered [36]. Regardless the poor data quality described for Brazilian public databases [29], the expected impairment in sensitivity was clearly not the case herein.

While Brazilian HIV/AIDS surveillance system greatly relies on linking government databases, the accuracy of procedure critically determines the validity of the final merged data [14]. An accurate algorithm would potentially improve the surveillance by reducing the underreporting and aiding to gather information in a timely fashion [43]. By combining the death cases reported to the surveillance database to those identified through a reliable record linkage algorithm, it will be possible to get better estimates on the mortality rate for both case surveillance and research purposes.

By using a bona fide gold standard to ascertain vital status information, we were able to compare two distinct algorithms, which proved to be accurate enough in retrieving vital status information from HIV/AIDS patients in the official mortality database. RlS had a consideqrably lesser runtime but intrinsically required manually reviewing a considerable proportion of the matched registries. On the other hand, PLA spared manual reviewing at no expense of accuracy. Choosing which is more suitable depends on a clear trade-off between processing time and the structure for a supervised procedure. Improvements in computation times are warranted should PLA be used in larger datasets on a routine basis, though unarguably both algorithms met the criteria to that end.

5. Conclusions

PLA and RlS are two free algorithms accurate enough in retrieving the vital status of people living with HIV/AIDS from the Brazilian mortality database. The routines had overlapping accuracies when good quality data were entered, though PLA presented higher sensibility when data lacked quality, which resulted in a significantly higher positive predictive value. RlS spent less processing time but intrinsically required manually inspecting a large proportion of the matched registries, whereas PLA spared manual reviewing at no expense of accuracy.

Authors' contributions

All authors have made substantial contributions to the conception and design of the study and to the analysis and interpretation of the data. All authors approved the final version of the manuscript.

Conflicts of interest

None.

Funding

This work was supported by the Carlos Chagas Filho Research Support Foundation/FAPERJ (grant number E-26/201.471/2014), the Strategic Health Research Support Program/PAPES (grant number 407446/2012-5) and the National Council for Scientific and Technological Development/CNPq (grant numbers 407446/2012-5 and 304614/2012-2). The study protocol has been approved by the local institutional review board (approval number 1172797).

Summary Points

What was already known on the topic:

- A number of record linkage algorithms are currently available to gather longitudinal information for entire populations.

- HIV/AIDS monitoring systems are rather depending on cross-referencing information from people living with HIV/AIDS (PLWHA) to mortality databases, thus accurate and computationally feasible linkage strategies are critically needed.

What this study added to our knowledge:

- Python language algorithm and Reclink software are two free algorithms which proved to be accurate enough in retrieving vital information from PLWHA registered in public databases to the official mortality system in Brazil.
- The choice of which is the most suitable or even a combined application depends upon a trade-off between an increased processing time in a fully unsupervised process as for Python Language Algorithm over an optimized processing time spent in an interactive supervised algorithm as Reclink.

Acknowledgements

Nominally identified databases were kindly provided by the Department of STDs, AIDS and Viral Hepatitis from the Brazilian Health Ministry.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ijmedinf.2018.03.005.

References

[1] E.A. Sauleau, J.-P. Paumier, A. Buemi, Medical record linkage in health information systems by approximate string matching and clustering, BMC Med. Inform. Decis. Mak. 5 (2005) 32, http://dx.doi.org/10.1186/1472-6947-5-32.

[2] H.B. Newcombe, J.M. Kennedy, Record linkage: making maximum use of the discriminating power of identifying information, Commun. ACM 5 (1962) 563–566, http://dx.doi.org/10.1145/368996.369026.

[3] S.B. Dusetzina, S. Tyree, A.-M. Meyer, A. Meyer, L. Green, W.R. Carpenter, Linking Data for Health Services Research, Agency for Healthcare Research and Quality, US, 2014 http://www.ncbi.nlm.nih.gov/pubmed/25392892 (Accessed March 23, 2017).

[4] D.E. Clark, Practical introduction to record linkage for injury research, Inj. Prev. 10 (2004) 186–191 http://www.ncbi.nlm.nih.gov/pubmed/15178677 (accessed April 10, 2017).

[5] P. Christen, Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Springer, 2012.

[6] A. Waldenburger, D. Nasseh, J. Stausberg, Detecting duplicates at hospital admission: comparison of deterministic and probabilistic record linkage, Stud. Health Technol. Inform. 226 (2016) 135–138 http://www.ncbi.nlm.nih.gov/pubmed/27350486.

[7] Y. Zhu, Y. Matsuyama, Y. Ohashi, S. Setoguchi, When to conduct probabilistic linkage vs. deterministic linkage? A simulation study, J. Biomed. Inform. 56 (2015) 80–86, http://dx.doi.org/10.1016/j.jbi.2015.05.012.

[8] T. Mi, S. Rajasekaran, R. Aseltine, Efficient algorithms for fast integration on large data sets from multiple sources, BMC Med. Inform. Decis. Mak. 12 (2012) 59, http://dx.doi.org/10.1186/1472-6947-12-59.

[9] B. Li, H. Quan, A. Fong, M. Lu, Assessing record linkage between health care and Vital Statistics databases using deterministic methods, BMC Health Serv. Res. 6 (2006) 48, http://dx.doi.org/10.1186/1472-6963-6-48.

[10] E.I. Benchimol, L. Smeeth, A. Guttmann, K. Harron, D. Moher, I. Petersen, H.T. Sørensen, E. von Elm, S.M. Langan, R.W. RECORD Working Committee, The REporting of studies conducted using observational routinely-collected health data (RECORD) statement, PLoS Med. 12 (2015) e1001885, http://dx.doi.org/10.1371/journal.pmed.1001885.

[11] DIAHV, Boletim Epidemiológico HIV/AIDS, Brasília/Brasil, (2017) http://www.aids.gov.br/pt-br/pub/2017/boletim-epidemiologico-hivaids-2017.

[12] DIAHV, Critérios de definição de casos de AIDS em adultos e crianças, Ministério da Saúde, Brasília/Brasil, 2004http://www.aids.gov.br/sites/default/files/criterios_aids_2004.pdf.

[13] H. Brenner, I. Schmidtmann, C. Stegmaier, Effects of record linkage errors on registry-based follow-up studies, Stat. Med. 16 (1997) 2633–2643 http://www.ncbi.nlm.nih.gov/pubmed/9421866.

[14] M.G.P. Fonseca, C.M. Coeli, F. de Fátima de Araújo Lucena, V.G. Lucena, M.S. Carvalho, Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database, Cad. Saude Publica 26 (7) (2010) 1431–1438 http://www.ncbi.nlm.nih.gov/pubmed/20694369.

[15] A.E. Miranda, R. Dietze, E.L. Maciel, T.N. Prado, A.L. Caus, M.M. Silva, J.E. Golub, Tuberculosis and AIDS co-morbidity in children: linkage of databases from espírito Santo State, Brazil, J. Trop. Pediatr. 57 (2011) 296–298, http://dx.doi.org/10.1093/tropej/fmq087.

[16] A.E. Miranda, J.E. Golub, F. de F. Lucena, E.N. Maciel, M. de F. Gurgel, R. Dietze, Tuberculosis and AIDS co-morbidity in Brazil: linkage of the tuberculosis and AIDS databases, Braz. J. Infect. Dis. 13 (2009) 137–141, http://dx.doi.org/10.1590/S1413-86702009000200013.

[17] A.G. Pacheco, V. Saraceni, S.H. Tuboi, L.M. Lauria, L.H. Moulton, J.C. Faulhaber, B. King, J.E. Golub, B. Durovni, S. Cavalcante, L.H. Harrison, R.E. Chaisson, M. Schechter, Estimating the extent of underreporting of mortality among HIV-infected individuals in Rio de Janeiro, Brazil, AIDS Res. Hum. Retrovir. 27 (2011) 25–28, http://dx.doi.org/10.1089/aid.2010.0089.

[18] A.G. Pacheco, V. Saraceni, S.H. Tuboi, L.H. Moulton, R.E. Chaisson, S.C. Cavalcante, B. Durovni, J.C. Faulhaber, J.E. Golub, B. King, M. Schechter, L.H. Harrison, Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in Brazil, Am. J. Epidemiol. 168 (2008) 1326–1332, http://dx.doi.org/10.1093/aje/kwn249.

[19] V. Saraceni, S. Cohn, S.C. Cavalcante, A.G. Pacheco, L.H. Moulton, R.E. Chaisson, B. Durovni, J.E. Golub, Prevalent tuberculosis at HIV diagnosis in Rio de Janeiro, Brazil: the TB/HIV in Rio (THRio) Cohort, J. Acquir. Immune Defic. Syndr. 67 (2014) 98–101, http://dx.doi.org/10.1097/QAI.0000000000000247.

[20] A.G. Pacheco, B. Durovni, S.C. Cavalcante, L.M. Lauria, R.D. Moore, L.H. Moulton, R.E. Chaisson, J.E. Golub, AIDS-related tuberculosis in Rio de Janeiro, Brazil, PLoS One 3 (2008) e3132, http://dx.doi.org/10.1371/journal.pone.0003132.

[21] B. Grinsztejn, P.M. Luz, A.G. Pacheco, D.V. Santos, L. Velasque, R.I. Moreira, M.R. Guimaraes, E.P. Nunes, A.S. Lemos, S.R. Ribeiro, D.P. Campos, M.A. Vitoria, V.G. Veloso, Changing mortality profile among HIV-infected patients in Rio de Janeiro, Brazil: shifting from AIDS to non-AIDS related conditions in the HAART era, PLoS One 8 (2013) e59768, http://dx.doi.org/10.1371/journal.pone.0059768.

[22] A.G. Pacheco, S.H. Tuboi, S.B. May, L.F. Moreira, L. Ramadas, E.P. Nunes, M. Mercon, J.C. Faulhaber, L.H. Harrison, M. Schechter, Temporal changes in causes of death among HIV-infected patients in the HAART era in Rio de Janeiro, Brazil, J. Acquir. Immune Defic. Syndr. 51 (2009) 624–630, http://dx.doi.org/10.1097/QAI.0b013e3181a4ecf5.

[23] M. da S. do Brasil, Sistema de Controle Logístico de Medicamentos (Siclom), (2015) http://www.aids.gov.br/pagina/sistema-de-controle-logistico-de-medicamentos-siclom.

[24] M. da S. do Brasil, Sistema de Controle de Exames Laboratoriais da Rede Nacional de Contagem de Linfócitos CD4 + /CD8 + e Carga Viral (Siscel), (2015) http://www.aids.gov.br/pagina/2010/sistema-de-controle-de-exames-laboratoriais-da-rede-nacional-de-contagem-de-linfocitos-c (Accessed 7 June 2015).

[25] X.-H. Zhou, D.K. McClish, N.A. Obuchowski, Electronic Book Collection, Wiley InterScience (Online Service), Statistical Methods in Diagnostic Medicine, Wiley, 2011.

[26] K.R. Camargo, C.M. Coeli, Reclink: an application for database linkage implementing the probabilistic record linkage method, Cad. Saude Publica 16 (2000) 439–447, http://dx.doi.org/10.1590/S0102-311x2000000200014.

[27] S.H.A. da Silva, J.C. da Mota, R.S. da Silva, M.R. Campos, J.M. de A. Schramm, Descrição dos registros repetidos no Sistema de Informação de Agravos de Notificação, Brasil, 2008–2009, Epidemiol. E Serviços Saúde 25 (2016) 487–498, http://dx.doi.org/10.5123/S1679-49742016000300005.

[28] E.V. Rossetto, E.J. de A. Luna, Database linkage for surveillance of the influenza A(H1N1)pdm09 pandemic in Brazil, 2009–2010, Cad. Saude Publica 32 (2016), http://dx.doi.org/10.1590/0102-311X00014115.

[29] V.R. Girianelli, L.C.S. Thuler, G.A. e Silva, Quality of cervical cancer data system in the state of Rio de Janeiro, Southeastern Brazil, Rev. Saude Publica 43 (2009) 580–588.

[30] P.V. Guimarães, C.M. Coeli, R.C.A. Cardoso, R. de A. Medronho, S.C. Fonseca, R.S. Pinheiro, Reliability of data from a very low birth weight population in the Live Birth Information System 2005–2006, Rev. Bras. Epidemiol. 15 (2012) 694–704 http://www.ncbi.nlm.nih.gov/pubmed/23515766 (Accessed March 21, 2017).

[31] C.M. Coeli, K.R. de Camargo Jr., Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros, (2002), http://dx.doi.org/10.1590/S1415-790x2002000200006.

[32] S.B. Oliveira, Estudo de coinfecção de HIV/AIDS com as hepatites virais B e C no Brasil no período de 1999 a 2010, Universidade de Brasília, 2013, http://repositorio.unb.br/bitstream/10482/13600/1/2013_SilvanoBarbosadeOliveira.pdf.

[33] P. Braitstein, M.W.G. Brinkhof, F. Dabis, M. Schechter, A. Boulle, P. Miotti, R. Wood, C. Laurent, E. Sprinz, C. Seyler, D.R. Bangsberg, E. Balestre, J.A.C. Sterne, M. May, M. Egger, Antiretroviral Therapy in Lower Income Countries (ART-LINC) Collaboration, ART Cohort Collaboration (ART-CC) groups, Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries, Lancet 367 (2006) 817–824, http://dx.doi.org/10.1016/S0140-6736(06)68337-2.

[34] C.J. Machado, K. Hill, Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem, Cad. Saude Publica 20 (2004) 915–925, http://dx.doi.org/10.1590/S0102-311x2004000400005.

[35] S. Gomatam, R. Carter, M. Ariet, G. Mitchell, An empirical comparison of record linkage procedures, Stat. Med. 21 (2002) 1485–1496, http://dx.doi.org/10.1002/sim.1147.

[36] S.J. Grannis, J.M. Overhage, S. Hui, C.J. McDonald, Analysis of a probabilistic

record linkage technique without human review, AMIA Annu Symp Proc. (2003) 259–263 http://www.ncbi.nlm.nih.gov/pubmed/14728174 (Accessed 11 April 2017).

[37] C.W. Kabudula, B.D. Clark, F.X. Gómez-Olivé, S. Tollman, J. Menken, G. Reniers, The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa, BMC Med. Res. Methodol. 14 (2014) 71, http://dx.doi.org/10.1186/1471-2288-14-71.

[38] P. Christen, Data Matching, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, http://dx.doi.org/10.1007/978-3-642-31164-2.

[39] F. Nakhaee, A. McDonald, D. Black, M. Law, A feasible method for linkage studies avoiding clerical review: linkage of the national HIV/AIDS surveillance databases with the National Death Index in Australia, Aust. N. Z. J. Public Health 31 (2007) 308–312 http://www.ncbi.nlm.nih.gov/pubmed/17725006 (Accessed 16 May 2017).

[40] J.P. Machado, M. Martins, Leite I. da C, J.P. Machado, M. Martins, Leite I. da C, Qualidade das bases de dados hospitalares no Brasil: alguns elementos, Rev. Bras. Epidemiol. 19 (2016) 567–581, http://dx.doi.org/10.1590/1980-5497201600030008.

[41] E.H. Porter, W.E. Winkler, Approximate string comparison and its effect on an advanced record linkage system, Adv. Rec. Link. Syst. U. S. Bur. CENSUS, Res. Rep. (1997) 190–199 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.7347 Accessed April 11, 2017.

[42] M.A. Bohensky, D. Jolley, V. Sundararajan, S. Evans, D.V. Pilcher, I. Scott, C.A. Brand, Data Linkage A powerful research tool with potential problems, BMC Health Serv. Res. 10 (2010) 346, http://dx.doi.org/10.1186/1472-6963-10-346.

[43] C.M. Coeli, K. Rochel De Camargo, Relacionamento de Bases de Dados em Saúde, Cad. Saúde Coletiva. XiV, (2006), pp. 305–312 http://www.cadernos.iesc.ufrj.br/cadernos/images/csc/2006_2/artigos/francisca_fatima_2006_2.pdf (Accessed 29 March 2017).