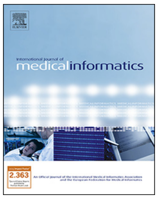
International Journal of Medical Informatics 114 (2018) 45–51

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

1386-5056/ © 2018 Elsevier B.V. All rights reserved.

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 nonmatches, 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.

Download English Version:

https://daneshyari.com/en/article/6926249

Download Persian Version:

https://daneshyari.com/article/6926249

Daneshyari.com