American Journal of Epidemiology
Vol. 168, No. 11
DOI: 10.1093/aje/kwn249
Advance Access publication October 9, 2008

Practice of Epidemiology

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

Antonio G. Pacheco, Valeria Saraceni, Suely H. Tuboi, Lawrence H. Moulton, Richard E. Chaisson, Solange C. Cavalcante, Betina Durovni, José C. Faulhaber, Jonathan E. Golub, Bonnie King, Mauro Schechter, and Lee H. Harrison

Initially submitted March 11, 2008; accepted for publication July 21, 2008.

Loss to follow-up is a major source of bias in cohorts of patients with human immunodeficiency virus (HIV) and could lead to underestimation of mortality. The authors developed a hierarchical deterministic linkage algorithm to be used primarily with cohorts of HIV-infected persons to recover vital status information for patients lost to follow-up. Data from patients known to be deceased in 2 cohorts in Rio de Janeiro, Brazil, and data from the Rio de Janeiro State mortality database for 1999–2006 were used to validate the algorithm. A fully automated procedure yielded a sensitivity of 92.9% and specificity of 100% when no information was missing. When the automated procedure was combined with clerical review, in a scenario of 5% death prevalence and 20% missing mothers' names, sensitivity reached 96.5% and specificity 100%. In a practical application, the algorithm significantly increased death rates and decreased the rate of loss to follow-up in the cohorts. The finding that 23.9% of matched records did not give HIV or acquired immunodeficiency syndrome as the cause of death reinforces the need to search all-cause mortality databases and alerts for possible underestimation of death rates. These results indicate that the algorithm is accurate enough to recover vital status information on patients lost to follow-up in cohort studies.

cohort studies; data collection; HIV; medical record linkage; mortality; software validation

Abbreviations: AIDS, acquired immunodeficiency syndrome; CI, confidence interval; HIV, human immunodeficiency virus; ICD-10, *International Classification of Diseases*, Tenth Revision; NPV, negative predictive value; PPV, positive predictive value; THRio, TB-HIV in Rio.

Database linkage is the process of comparing records from different databases that contain enough information to determine whether those records refer to the same person or, more generally, to the same entity (1).

There are 3 main types of record linkage: manual (or clerical), deterministic, and probabilistic. These methods can be combined, depending on the strategy used. The first type consists of manually comparing records in 2 databases and deciding whether they are true matches or not. This was the standard method used before the availability of computers. It is often highly labor-intensive and is sometimes not feasible, particularly when the amount of data is too large. Deterministic methods are classically based on exact-match comparisons of either 1 unique identifier common to both databases (e.g., Social Security number) or a combination of variables that allow unique discrimination (e.g., name, surname, date of birth, gender) (2–4). Probabilistic methods are also based on several variables, but comparisons are made on the basis of the prior probability of 2 records' belonging to the same entity and then calculating a maximum likelihood estimator to reach a score for similarity between records (1, 5, 6). The method (or combination

Correspondence to Dr. Antonio Guilherme Fonseca Pacheco, Programa de Computação Científica, Fundação Oswaldo Cruz, Avenida Brasil, 4365, Manguinhos, 21045-360, Rio de Janeiro, Brazil (e-mail: apacheco@fiocruz.br).

of methods) to be chosen depends on the type of analysis to be carried out with the linked data and the types of databases available (7).

Record linkage is widely used in population-based studies to make inferences about specific outcomes and in cohort studies to make inferences at the level of the individual (3, 7–10). Morbidity and mortality databases are often employed for this purpose, given their wide availability and the fact that their records generally contain sufficient information for linkage with other databases.

Investigators in cohort studies usually use linkage techniques to gather additional information about patients being followed over time. Even when studying cohorts with active follow-up, investigators tend to complement their information with external databases in order to minimize underreporting of specific conditions (e.g., vital status), including in their protocol a passive follow-up component. In the case of cohorts of human immunodeficiency virus (HIV)-infected patients, morbidity databases (e.g., tuberculosis, cancer) are important sources of additional information (11–13).

In Brazil, official surveillance and mortality databases contain variables, such as full name, date of birth, and mother's name (either maiden or married surname), that are suitable for linkage procedures because of their potentially high discriminatory power, particularly when used in combination.

In the present study, we describe the validation of a new deterministic linkage algorithm that we developed to be used for passive data collection with cohorts of HIV-infected patients. The algorithm has a hierarchical structure and allows for specific errors in names and dates of birth. It can be used in combination with clerical review of records that are not classified as true matches or are not excluded as nonmatches. Our main objectives when developing the algorithm were to maximize accuracy and to minimize the need for clerical review.

MATERIALS AND METHODS

Data sources

Three data sources were used in this study. The Rio de Janeiro cohort database was originally designed to validate the World Health Organization HIV staging system in a developing country (14). It currently comprises information from 2,666 HIV-infected patients being followed at the Clementino Fraga Filho University Hospital in Rio de Janeiro, Brazil. All patients are aged 16 years or older and are included only if they have made at least 1 follow-up visit. The rate of loss to follow-up between 2000 and 2005 in the Rio de Janeiro cohort was 2.9 per 100 person-years.

The TB-HIV in Rio (THRio) Study is an ongoing cohort study designed to assess the impact of implementing isoniazid prophylactic therapy among HIV-positive patients with indications for prophylaxis in Rio de Janeiro. It has enrolled more than 15,000 patients from 29 clinics, where care is provided both for HIV and for tuberculosis (15, 16). There has not yet been enough follow-up time to calculate accurately the rates of loss to follow-up in this study.

The third database is the Rio de Janeiro State mortality database for 2000–2006, with a total of 835,066 records. The Rio de Janeiro State mortality database is part of the ''Sistema de Informação sobre Mortalidade'' database, which is the official mortality system in Brazil. The death certificate is a standardized form that is filled out by a physician. It includes demographic information and primary, secondary, and contributing causes of death coded according to the *International Classification of Diseases*, Tenth Revision (ICD-10), among other variables. An electronic version of these forms was introduced countrywide in 1979. Information that can identify patients, such as name, mother's name, date of birth, and address, is also recorded and was made available through a special request to the state health department. According to the Brazilian Ministry of Health, the mortality system in Rio de Janeiro State has 100% coverage of deaths (17), even though the percentage of undefined causes of death remained somewhat high (9.3%) in 2005.

Data linkage between both cohort databases and the mortality database is part of routine procedures for assessment of vital status among patients who are lost to follow-up and was approved by the institutional review boards of all involved institutions. Data from patients known to be deceased through an independent source (generally medical charts) with identifying variables were used for validation purposes.

To validate the algorithm, we assembled test data sets and then linked them with the mortality database. We then studied the outcome ''finding a record in the mortality database.'' To determine the sensitivity of the linkage to the mortality database for identifying deceased patients, all patients known to be deceased and who had complete information on full name, date of birth, and mother's name (either maiden or married surname) in the Rio de Janeiro cohort between 2000 and 2005 (53 patients) and in the THRio cohort between 2003 and 2006 (315 patients) were included in the analysis.

To assess specificity, we incorporated into the test database records that were not supposed to be in the state mortality database between 2000 and 2006 and that would be subject to similar typing mistakes as those for the patients known to be deceased. We chose to use a random sample of control records of patients who died in 1999, a year that was not included in the linkage.

The overall completeness of information for the THRio cohort was 98.3% for full information and 99.7% for name and date of birth. In the Rio de Janeiro cohort, 60% had full information and 100% had at least name and date of birth.

Data preprocessing

The first step in data linkage was preprocessing of data to guarantee that all variables conformed to the same format. For names, all letters were capitalized, and accents and characters other than letters were removed. Suffixes referring to a person of the same name, such as the individual's father (Junior, Filho, etc.), were also removed. A specific software function (see supplementary data posted on the *Journal*'s website (http://aje.oxfordjournals.org/)) was

developed for this purpose and has the ability to preprocess a string field as a whole, either with Windows-based Latin alphabet encoding (cp1252) or with the DOS-based alphabet, which is still used in older ''.dbf'' files (cp850).

Linkage algorithm

To avoid exponential growth of processing time, we first blocked records (i.e., grouped them) by means of a phonetic code adapted from the original Soundex algorithm (18) to account for Brazilian Portuguese names (see supplementary data for details (http://aje.oxfordjournals.org/)). Blocks were composed by combining either the phonetic codes from the first and last names, the phonetic codes from the mother's first and last names, or the phonetic codes from the first name and the mother's first name. We used the third category to account for last names that are difficult to spell and that are recorded equally for the individual and his/her mother in both databases but are misspelled in one of them.

Records within each block were then compared, using exact comparisons and also allowing for some errors, in a hierarchical fashion, as described below. Errors in name fields were evaluated by means of the phonetic codes and also by a string similarity score, based on a recursive longer common substring algorithm, implemented in the ''difflib'' library from the programming language Python (19). Dates of birth were allowed to have, at most, a 1-digit mistake in any position or the common swap between day and month (only if they were exactly the same, but swapped). Score values used in the algorithm as described in Table 1 were chosen empirically in the beginning of the algorithm development, using different data sources (municipal databases for surveillance of acquired immunodeficiency syndrome (AIDS) and tuberculosis; data not shown). The combinations of these measurements and the values for the scores determine several levels of inclusion—which in the present paper are referred to as ''automatic codes'' and depend on how much information is available, as shown in Table 1.

Records with complete information (automatic codes 0–7; Table 1) are treated independently from records with missing data (automatic codes 8–10 when mother's name is missing; Table 1). Whenever a pair of records is neither automatically included with one of the inclusion codes described nor automatically excluded by the criteria in Table 1, this pair is kept in the final merged database, marked as an unresolved pair for possible further clerical review. The algorithm is hierarchical in the sense that lower codes mean more similar records—0 and 8 are perfect matches, but codes 0–7 are used for records with full information and thus are more robust than codes 8–10 for records that are missing mother's name. The algorithm is not ''greedy'' in that the same record in the test database linked with a 0 code to one record could also be linked to another one with a code 7, for example. This feature is important, because the algorithm can also be used for databases with 1-to-many relations, as in the case of tuberculosis surveillance databases. For mortality, which is supposed to have a 1-to-1 relation with the cohort databases, multiple matches for the same patient can easily be resolved by automatically picking the match with the lowest value. This was done in the present study.

If a pair was neither included nor excluded, it was eligible for clerical review. For records with name only, only perfect matches were considered.

The algorithm was written in Python for Windows (19).

Algorithm validation

We used 3 different scenarios to validate the algorithm. First we considered a hypothetical situation in which patients lost to follow-up in a cohort of HIV-infected persons would be searched for in the mortality database, and we assumed that 50% of these lost patients had actually died. Thus, we constructed a database by combining the 368 records of patients known to be deceased in the cohorts with a random sample of 368 records from the 1999 mortality database. In this scenario, we compared accuracy for exact matches between the records in all fields with the automatic inclusion codes, when 1) full information was available for all individuals, 2) only name and date of birth were available, and 3) only name was available.

In the second scenario, we tested the impact on accuracy if all patients in a cohort of HIV-infected persons were linked to the mortality database, considering that only 5% of the patients were truly deceased. This is a reasonable percentage for an open cohort of HIV-infected patients in developing countries, where death rates are generally around 5 per 100 person-years (20). In this case, we made up a data set with 368 records of patients known to be deceased in the cohorts and by randomly selecting 6,992 patients ($368/0.05 - 368 = 6{,}992$) from the 1999 mortality database. In these 2 scenarios, ties were resolved automatically by choosing the pair with the lowest score, and no manual search was performed—the aim being merely to assess the potential impact on accuracy of missing information in the test database.

In the third scenario, we mimicked a situation similar to what one may encounter in practical research with a cohort of HIV-infected patients: assuming a 50% prevalence of deaths among patients lost to follow-up and a 20% prevalence of records missing the mother's name. The database was set up for this scenario in the same way as it was for the first scenario, but 20% of mothers' names were randomly deleted from the test database. In this run, we did not consider records that were missing date of birth. Unresolved pairs were submitted to clerical review by 2 independent researchers, and disagreements were resolved by a third reviewer. In addition, records with automatic inclusion codes were manually reviewed for quality control purposes. To minimize selection bias, reviewers had no access to the group membership status of records being reviewed.

Sensitivity, specificity, positive predictive values (PPVs), and negative predictive values (NPVs) were calculated for the experiments, along with 95% confidence intervals, using appropriate methods (21). The total numbers of records in the test databases were used as denominators for calculations.

For records of patients found in the mortality system, we assessed the proportion of cases for which AIDS-related ICD-10 codes (codes B20–B24) were not mentioned on the death certificate. The coding system used for death

Table 1. Classification of Matched Records Used to Validate a Record-Linkage Algorithm, Brazil, 1999–2006

Automatic Inclusion Codes[a]	Patient's Name	Date of Birth	Mother's Name[b]
0	Exact	Exact	Exact
1	Exact	Exact	Same PC
2	Exact	1 error or swap	Exact
3	Exact	1 error or swap	Same PC
4	Score $>$ 0.75	Exact	Exact
5	Score $>$ 0.75	1 error or swap	Exact
6	Score $>$ 0.75	Exact	Same PC + score $>$ 0.75
7	Score $>$ 0.9	1 error or swap	Score $>$ 0.8
8	Exact	Exact	Missing
9	Exact	1 error or swap	Missing
10	Score $>$ 0.9	Exact	Missing
Exclusion[c]	Not missing	$>$1 error	Different PC
	Score $\leq$ 0.9	$>$1 error	Score $\leq$ 0.8
	Not missing	$>$1 error	Score $\leq$ 0.7
	Score $<$ 0.8	Not missing	Not missing
	Not missing	Day, month, and year are different	Missing
	Score $<$ 0.8		Missing

Abbreviation: PC, phonetic code.

[a] After passing the first blocking phase: same PC of patient's first and last name OR same PC of mother's first and last name OR same PC of patient's and mother's first names.

[b] PC is for mother's name only in this case.

[c] Records that are not included or excluded are left over for clerical review. Score values were chosen empirically (see text).

certificates follows the World Health Organization guidelines (22).

For a preliminary practical application of the algorithm, rates of loss to follow-up were compared before and after the algorithm was used for the Rio de Janeiro cohort and death rates were compared before and after the algorithm was used for both cohorts. Exact Poisson 95% confidence intervals are presented for the differences. Calculations were done in the R software environment (23).

RESULTS

Table 2 shows results from the first 2 scenarios for assessing the impact of missing information on the accuracy of the algorithm. As expected, sensitivity for exact matches increased when less information was available for linking the records, while the addition of the automatic codes without manual review represented a significant increase both when full information was available (from 50.8% to 92.9%) and when the mother's name was missing (from 71.2% to 91.8%). Specificity for both cases was very high, and no misclassification was made by the algorithm when the death prevalence was 50% (PPV = 100%). To evaluate the impact of having 5% or 50% of prevalence, we compared PPV and NPV in these 2 scenarios. While PPV was 100% at 50% prevalence and no misclassifications occurred, it was reduced, as expected, at 5% prevalence. Even though the percentages of automatic codes with full information and missing mother's name were still high (99.4% and 92.6%, respectively), these represented 2 false-positive cases in the first instance and 27 in the second. Accuracy for records with patients' names only was not as good as with the other variables, even when we considered exact matches only (Table 2, last column), since sensitivities were lower than the ones for automatic codes for the other scenarios. Specificity in this case was very low, with a PPV of only 81.2% in the 50% scenario, yielding 66 false-positive cases, and as low as 19.5% in the 5% scenario, reaching 1,175 false-positive cases.

With reference to records to be manually checked, 1,189 of those with full information and 4,146 of those with missing mother's name would have to be searched for a prevalence of 50% and 9,351 and 48,333, respectively, would have to be searched for a prevalence of 5%.

In the third scenario, with 50% prevalence and 20% of the records missing mother's name, the results obtained were: sensitivity = 96.5% (95% confidence interval (CI): 94.0, 98.1); specificity = 100% (95% CI: 99, 100); PPV = 100% (95% CI: 99.0, 100); and NPV = 96.6% (95% CI: 94.2, 98.2). Manual review was performed on the 1,929 pairs that the algorithm was not able to include or exclude

Table 2. Accuracy of Exact Matches and Automatic Codes When Records in the Test Database Have Full or Partial Information (50% and 5% Prevalence Scenarios), Brazil, 1999–2006

Accuracy	Full Information				No Mother's Name				Name Only[a] (Exact Match[b])	
	Exact Match[b]		Automatic Codes[c]		Exact Match[b]		Automatic Codes[c]			
	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Sensitivity	50.8	45.6, 56.0	92.9	88.3, 94.2	71.2	66.3, 75.8	91.8	88.6, 94.4	77.4	72.8, 81.6
Specificity	100.0	99.0, 100.0	100.0	99, 100.0	100.0	99.0, 100.0	100.0	99.0, 100.0	82.1	77.8, 85.8
50% prevalence										
PPV	100.0	98.0, 100.0	100.0	98.9, 100.0	100.0	98.6, 100.0	100.0	98.9, 100.0	81.2	76.7, 85.1
NPV	67.0	62.9, 70.9	93.4	90.5, 95.6	77.6	73.6, 81.3	92.5	89.4, 94.9	78.4	74.0, 82.4
5% prevalence										
PPV	100.0	98.0, 100.0	99.4	97.9, 99.9	98.9	96.7, 99.8	92.6	89.4, 95.0	19.5	17.5, 21.6
NPV	97.5	97.1, 97.8	99.6	99.5, 99.8	98.5	98.2, 98.8	99.6	99.4, 99.7	98.6	98.3, 98.9

Abbreviations: CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value.
[a] Since only name was available in this case, only exact matches were considered.
[b] Exact match means a perfect match between the available variables in both databases.
[c] The automatic inclusion codes listed in Table 1.

as a true match. Of those, 9 pairs were considered true matches by reviewer 1 and 11 were considered true matches by reviewer 2. The 2 disagreements were submitted to a third reviewer, who considered 1 of them a true match. In a manual review of the automatic codes, all of them were considered true matches.

The combination of automatic codes and clerical review yielded high sensitivity and specificity. For this test database, the PPV was 100% and the NPV was 96.6%.

Among the 355 patients who were found by the algorithm, 85 (23.9%) did not have HIV- or AIDS-related ICD-10 codes (codes B20–B24) given as the underlying cause of death.

Before the algorithm was used, the rate of loss to follow-up in the Rio de Janeiro cohort between 2000 and 2005 was 2.9 per 100 person-years; it dropped to 2.1 per 100 person-years after recovery from the mortality system (difference = −0.8, 95% CI: −1.1, −0.6). In the same period, the mortalxity rate increased with inclusion of deaths from the mortality system, from 2.2 per 100 person-years to 3.2 per 100 person-years (difference = 1.0, 95% CI: 0.7, 1.3). For the THRio Study cohort, the death rate in 2006 before the use of the algorithm was 1.2 per 100 person-years; it increased to 4.2 per 100 person-years after deaths were recovered for all patients in the cohort, using automatic codes only, without manual review (difference = 3.0, 95% CI: 2.7, 3.3).

DISCUSSION

The deterministic algorithm validated in the present study was developed primarily to assist investigators actively following cohorts of HIV-infected persons to improve their performance by searching for patients lost to follow-up in mortality databases. The performance characteristics of the algorithm were excellent, with a sensitivity of over 90% for automatic codes, either in the 5% prevalence scenario or in the 50% prevalence scenario, which was minimally affected when mother's name was not available. These figures were well over the sensitivity for exact matches of 50% and 71% when full information was available and the mother's name was missing, respectively. Specificity was close to 100% for all cases, meaning that not a single pair was misclassified as a false-positive, but when we considered records with patients' names only, even for exact matches specificity was unacceptably low (approximately 82%). These results are in agreement with those of the study by Quantin et al. (24), who found that date of birth and first and last patient's name would have sufficient discriminatory power, even though their study was carried out using a probabilistic approach and they did not test mother's name as one of the variables. In the 5% scenario, the PPV remained close to 100% in all situations; there were 2 false-positives in the full information data set and an excess of 27 when mother's name was missing. This indicates that even though false-positives are very unlikely (PPV = 98.9% and PPV = 92.5%, respectively), caution must be taken when designating a patient deceased.

In the third scenario, with 50% prevalence and with 20% of the records missing mother's name, clerical review increased sensitivity to over 96%, while preserving 100% specificity.

Although sensitivity was high, it was still impossible to find 13 patients reported as deceased in medical charts. There are 2 possible explanations for this finding: 1) these patients indeed were not included in the mortality database or 2) these patients were in the database but the algorithm was not able to find their records. In the former case, if the patient was in fact still alive, he or she was truly lost to follow-up. On the other hand, if the patient was indeed deceased, either the event was not detected by the system or the patient had moved out of the state and died elsewhere. In the case of patients who were in the mortality database, the main reasons for not finding a record were major spelling errors—especially for the first letter, which is very sensitive to Soundex-like phonetic algorithms, but also deletion of the

last name in the case of persons with 4 or more names—and incorrectly entered dates of birth.

The number of records left for manual review suggests that the best option is to preselect records to be linked to the databases in order to increase the number of patients who could be found, decreasing clerical review.

Even though probabilistic algorithms have been extensively studied and there is at least 1 algorithm validated for Brazilian databases (25, 26), we chose to employ a deterministic approach, allowing for some uncertainty with regard to the variables used. This decision was based on the fact that even though probabilistic algorithms tend to yield higher sensitivities, they do so by sacrificing specificity—which is not a major problem when studying population-based characteristics, given that false-positives and false-negatives would tend to cancel out (7, 27). Conversely, for inferring the vital status of individual patients being followed in a cohort, this approach is not advisable and deterministic algorithms are more indicated (7), given that ethical problems may emerge when a patient is declared deceased and he or she shows up for a subsequent visit. In either case, caution should always be exercised, since bias due to false-positives and false-negatives would lead to over- or underestimation of the parameters being studied, although the impact of false-positives on overestimation tends to be more severe than the impact of false-negatives on underestimation (27).

One of the reasons cohorts with active tracing of patients might suffer from loss to follow-up is that information on deaths which occur in other health-care facilities might be outside the area of the clinic where routine care was provided, especially if the cause of death was not related to AIDS. In cohorts of patients who are intrinsically highly prone to morbidity and/or mortality, as with HIV-infected patients, this can be particularly problematic. For example, in a study involving 6,498 patients being followed in 18 treatment programs in lower-income countries, the estimated death rate 1 year after initiation of antiretroviral therapy would increase from 6.4 per 100 person-years to 15 per 100 person-years if mortality among those lost to follow-up was similar to that observed in patients without antiretroviral therapy (20). In another report on cohorts in sub-Saharan countries, 41% of patients had an unknown vital status on medical charts, and 65% of those, initially considered lost to follow-up, were found to be deceased after appropriate vital status investigation procedures were applied (28).

A practical application of the algorithm showed very good results for both cohorts. In the THRio Study cohort, applying the algorithm for all patients with the 2006 mortality database, there was a significant increase in the death rate for that year, even using automatic codes only. For the Rio de Janeiro cohort, the impact of the algorithm on patients lost to follow-up was significant both in reducing losses to follow-up and in increasing the death rate for the period 2000–2005.

Finally, the fact that almost 24% of the death certificates of cases that were found through our algorithm did not have AIDS-related ICD-10 codes (codes B20–B24) mentioned on them underscores the need to search in all-cause mortality databases and not to restrict searches to HIV/AIDS deaths. On the other hand, this finding suggests that official HIV/AIDS mortality figures that are based solely on the mortality system might significantly underestimate the true figures, a possibility that should be formally evaluated. This might lead to adjustments in mortality statistics in Brazil.

ACKNOWLEDGMENTS

Author affiliations: Escola Nacional de Saúde Pública Sérgio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil (Antonio G. Pacheco); Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil (Antonio G. Pacheco); Infectious Diseases Epidemiology Research Unit, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania (Antonio G. Pacheco, Suely H. Tuboi, Lee H. Harrison); School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania (Lee H. Harrison); Communicable Diseases Program, Rio de Janeiro Municipal Health Secretariat, Rio de Janeiro, Brazil (Valeria Saraceni, Solange C. Cavalcante, Betina Durovni); Projeto Praça Onze, Hospital Escola São Francisco de Assis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil (Suely H. Tuboi, José C. Faulhaber, Mauro Schechter); Department of International Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland (Lawrence H. Moulton); Center for Tuberculosis Research, School of Medicine, Johns Hopkins University, Baltimore, Maryland (Richard E. Chaisson, Jonathan E. Golub, Bonnie King); Instituto de Pesquisa Evandro Chagas–Fundação Oswaldo Cruz, Rio de Janeiro, Brazil (Solange C. Cavalcante); and AIDS Research Laboratory, Hospital Universitario Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil (Mauro Schechter).

This study was partially supported by the Fogarty International Center, US National Institutes of Health (grant 3 D43 TW01038 to the University of Pittsburgh); by Cooperative Agreement 1 U01 AI069923 from the National Institute of Allergy and Infectious Diseases and the National Cancer Institute, US National Institutes of Health; and by the Bill and Melinda Gates Foundation.

The authors thank Dr. Oswaldo G. Cruz for his contribution to the adapted Soundex algorithm.

Conflict of interest: none declared.

REFERENCES

1. Howe GR. Use of computerized record linkage in cohort studies. *Epidemiol Rev.* 1998;20(1):112–121.
2. Li B, Quan H, Fong A, et al. Assessing record linkage between health care and vital statistics databases using deterministic methods [electronic article]. *BMC Health Serv Res.* 2006;6:48.
3. Gomatam S, Carter R, Ariet M, et al. An empirical comparison of record linkage procedures. *Stat Med.* 2002;21(10):1485–1496.
4. Clark DE, Hahn DR. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. *Proc Annu Symp Comput Appl Med Care*. 1995:397–401.

5. Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc.* 1969;64(328):1183–1210.
6. Newcombe HB, Kennedy JM, Axford SJ, et al. Automatic linkage of vital records. *Science*. 1959;130(3381):954–959.
7. Clark DE. Practical introduction to record linkage for injury research. *Inj Prev.* 2004;10(3):186–191.
8. The West of Scotland Coronary Prevention Study Group. Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. *J Clin Epidemiol*. 1995; 48(12):1441–1452.
9. Roos LL, Walld R, Wajda A, et al. Record linkage strategies, outpatient procedures, and administrative data. *Med Care*. 1996;34(6):570–582.
10. Whiteman D, Murphy M, Hey K, et al. Reproductive factors, subfertility, and risk of neural tube defects: a case-control study based on the Oxford Record Linkage Study Register. *Am J Epidemiol*. 2000;152(9):823–828.
11. Ahmed AB, Abubakar I, Delpech V, et al. The growing impact of HIV infection on the epidemiology of tuberculosis in England and Wales: 1999–2003. *Thorax*. 2007;62(8):672–676.
12. Frisch M, Biggar RJ, Engels EA, et al. Association of cancer with AIDS-related immunosuppression in adults. *JAMA*. 2001;285(13):1736–1745.
13. Polesel J, Clifford GM, Rickenbach M, et al. Non-Hodgkin lymphoma incidence in the Swiss HIV Cohort Study before and after highly active antiretroviral therapy. *AIDS*. 2008; 22(2):301–306.
14. Schechter M, Zajdenverg R, Machado LL, et al. Predicting CD4 counts in HIV-infected Brazilian individuals: a model based on the World Health Organization staging system. *J Acquir Immune Defic Syndr.* 1994;7(2):163–168.
15. Golub JE, Saraceni V, Cavalcante SC, et al. The impact of antiretroviral therapy and isoniazid preventive therapy on tuberculosis incidence in HIV-infected patients in Rio de Janeiro, Brazil. *AIDS*. 2007;21(11):1441–1448.
16. Moulton LH, Golub JE, Durovni B, et al. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clin Trials*. 2007;4(2):190–199.
17. Ministério da Saúde, Secretaria de Vigilância em Saúde. *Sistema Nacional de Vigilância em Saúde: Relatório de Situação: Rio de Janeiro*. Brasília, Brazil: Ministério da Saúde; 2007. (Série C: Projetos, Programas e Relatórios).
18. National Archives and Records Administration. *The Soundex Indexing System*. College Park, MD: National Archives and Records Administration; 2007. (http://www.archives.gov/genealogy/census/soundex.html). (Accessed January 3, 2008).
19. Lutz M. *Programming Python*. 3rd ed. Sebastopol, CA: O'Reilly Media, Inc; 2006.
20. Braitstein P, Brinkhof MW, Dabis F, et al. Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries. *Lancet*. 2006;367(9513):817–824.
21. Zhou X-H, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. New York, NY: Wiley-Interscience; 2002.
22. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision. Vol 2. Instruction Manual*. 2nd ed. Geneva, Switzerland: World Health Organization; 2004.
23. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2006.
24. Quantin C, Binquet C, Bourquard K, et al. Which are the best identifiers for record linkage? *Med Inform Internet Med*. 2004; 29(3–4):221–227.
25. Camargo KR Jr, Coeli CM. RecLink: an application for database linkage implementing the probabilistic record linkage method [in Portuguese]. *Cad Saude Publica*. 2000;16(2): 439–447.
26. Coutinho ES, Coeli CM. Accuracy of the probabilistic record linkage methodology to ascertain deaths in survival studies [in Portuguese]. *Cad Saude Publica*. 2006;22(10):2249–2252.
27. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. *Stat Med*. 1997;16(23):2633–2643.
28. Anglaret X, Toure S, Gourvellec G, et al. Impact of vital status investigation procedures on estimates of survival in cohorts of HIV-infected patients from sub-Saharan Africa. *J Acquir Immune Defic Syndr.* 2004;35(3):320–323.