Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database

Acurácia da estratégia de relacionamento probabilístico em identificar óbitos entre casos de AIDS notificados no Sistema de Informação de Agravos de Notificação (SINAN)

Maria Goretti Pereira Fonseca [1,2]
Cláudia Medina Coeli [3]
Francisca de Fátima de Araújo Lucena [4]
Valdilea Gonçalves Veloso [2]
Marilia Sá Carvalho [5]

[1] Diretoria Regional de Brasília, Fundação Oswaldo Cruz, Brasília, Brasil.
[2] Instituto de Pesquisa Clínica Evandro Chagas, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.
[3] Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.
[4] Ministério do Desenvolvimento Social e Combate à Fome, Brasília, Brasil.
[5] Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.

Correspondence
M. G. P. Fonseca
Diretoria Regional de Brasília, Fundação Oswaldo Cruz.
SHIN QL 07, conjunto 06, casa 18, Brasília, DF 71515-065, Brasil.
gorettifonseca@fiocruz.br

Abstract

Since record linkage errors can bias measures of disease occurrence and association, it is important to assess their accuracy. The aim of this study is to assess the accuracy of a multiple pass probabilistic record linkage strategy to identify deaths among persons reported to the Brazilian AIDS surveillance database. An HIV/AIDS national surveillance database (N = 559,442) was linked to a total of 6,444,822 deaths registered (all causes) in the Brazilian mortality database. To estimate standard measures of accuracy, we selected all AIDS cases with a date of death registered in the surveillance database from 2002 to 2005 (N = 19,750) and 38,675 cases known to be alive in 2006. The linkage strategy presented a sensitivity of 87.6% (95%CI: 87.1-88.2), a specificity of 99.6% (95%CI: 99.6-99.7), and a positive predictive value of 99.2% (95%CI: 99.1-99.3). We observed a small variation in the validity measures according to some putative predictors of mortality. Our findings suggest that even large and heterogeneous databases can be linked with a satisfactory accuracy.

Medical Record Linkage; Information Systems; Acquired Immunodeficiency Syndrome; Mortality

Introduction

The Brazilian National AIDS Program has been acknowledged as a success in controlling the epidemic. Its major tools to support the epidemic control are based on prevention measures, surveillance case reporting, monitoring people living with HIV/AIDS through laboratory tests, and universal access to AIDS treatment for those in need [1]. That policy has generated three major electronic databases: SINAN-AIDS (Information System for Notifiable Diseases of AIDS Cases), SISCEL (Laboratory Test Control System) and SICLOM (System for Logistic Control of Drugs) [2]. Alongside these databases, there are a variety of health information systems that are available for surveillance concerning mortality, live births and ambulatory and hospital care funding by the Unified National Health System (SUS) in both public and private institutions [3].

Record linkage has been increasingly used in AIDS surveillance [2,4] and research [5,6,7,8]. In the Brazilian National AIDS Program, record linkage is carried out by the Surveillance Unit aiming to verify underreporting of cases and eliminate duplicated cases with improving results [9]. As a unique identifier is not available in the health databases, identification fields were used together and a probabilistic approach was adopted. Probabilistic record linkage is based on similar variables present in the databases to be linked (e.g.: name, sex, date of birth, area of residence).

These personal identifiers are used together in order to determine how likely a pair of records refers to the same individual [10]. The accuracy of the probabilistic linkage process is strongly dependent on the number and quality of the personal identifiers available to be compared, as well as the strategy adopted to link the databases [5,10]. Because record linkage errors can bias measures of disease occurrence and association [11,12,13], it is important to assess the accuracy of record linkage methods employed for surveillance and research purposes.

The aim of the present study was to assess the accuracy of a multiple pass probabilistic record linkage strategy to identify deaths among persons reported to the Brazilian AIDS surveillance database.

Methods

Data sources

SINAN-AIDS is the most important electronic AIDS case surveillance database in Brazil. The system is implemented in every municipality that is eligible to report AIDS cases to the state and federal levels, and it has been regularly updated. It registers all cases reported since 1980, with 506,499 AIDS cases up to June, 2008 [9], including underreported cases recovered, recording socio-demographic as well as epidemiological information. Brazil has adopted its own AIDS case definitions for surveillance purposes: the Brazilian CDC, where some diseases are presumptive but not definitive, besides the CD4 count below 350cells/mm^3; Rio de Janeiro/Caracas, a point bases case-definition for minor and major signs; and the death case definition, when a case is identified only through the death certificate [14]. The database is processed on a regular basis by the Surveillance Unit of the Brazilian National AIDS Program, applying a probabilistic record technique to eliminate duplicated records and to improve database completeness [9]. The SISCEL is a data system developed to monitor laboratory tests, such as lymphocytes CD4+ T cell counts and viral load tests, for people living with HIV and AIDS being followed in the public health sector. Implemented in 2002, by July 2006, 88 labs were using SISCEL to register CD4 test results and 75 to register viral load test results, covering 90% of all CD4 and viral load tests done by the public health sector (SISCEL. http://www.aids.gov.br/data/Pages/LUMIS61CDFF9FENIE.htm, accessed on 08/Aug/2009). By June 2007, the system registered the lab results of 220,000 HIV positive individuals. The SICLOM was also developed to control the logistic of AIDS treatment distribution and the system shares the same patient list with SISCEL. From 2002 to 2006, 133,768 patients were registered in SICLOM. The Brazilian Mortality Information System (SIM) registers all deaths, using a standard death certificate adopted throughout the entire country. The 10th Revision of the International Classification of Diseases (ICD-10) has been used since 1996.

We created a combined database that included both HIV and AIDS cases (N = 559,442 individuals) by linking SINAN-AIDS to SISCEL and SICLOM databases, applying the linkage strategy adopted by the Surveillance Unit of the Brazilian National AIDS Program [2], including all individuals in each database. This database was then linked to a total of 6,444,822 deaths registered (including both AIDS and other conditions as underlying cause of death) in the SIM from 2000 to 2006.

Record linkage strategy

Linkage was performed using RecLink III software [15]. The databases were preprocessed in order to achieve standardization and parsing of the fields that were selected to be used as matching and/or blocking variables. A three-pass blocking strategy was applied using different keys formed by the combination of the following fields: phonetics codes of first name and last name; sex; year of birth and code of municipality of residence. Name, mother's name and date of birth were used as matching fields with parameter estimates being obtained by the EM algorithm [10]. The field's name and mother's name were compared using the Levenshtein distance string comparator metric [16], whereas for the date of birth an exact (character-by-character) algorithm was used. For each link of records a composite weight was calculated with the sum of the agreement or the disagreement weight for each field being compared [10]. The scores ranged between -13.20 and +35.79. Links that presented a composite weight higher than 18.9 were designated true matches and those with a composite weight below 10.85 were considered false matches. Between 10.85 and 18.9 they were considered potential matches and were manually reviewed by one of the authors (F.F.A.L.).

Data analysis

To assess the accuracy of the strategy used to link the HIV/AIDS database to the mortality data, we selected all AIDS cases reported in SINAN up to June 2007 with date of diagnosis between 2002 and 2005 (N = 106,283). Cases diagnosed before

2002 were excluded because personal identifiers were not available in the mortality database for the entire country before 2002. Individuals registered only in SISCEL and/or in SICLOM were not analyzed because of a lack of information about vital status in these systems.

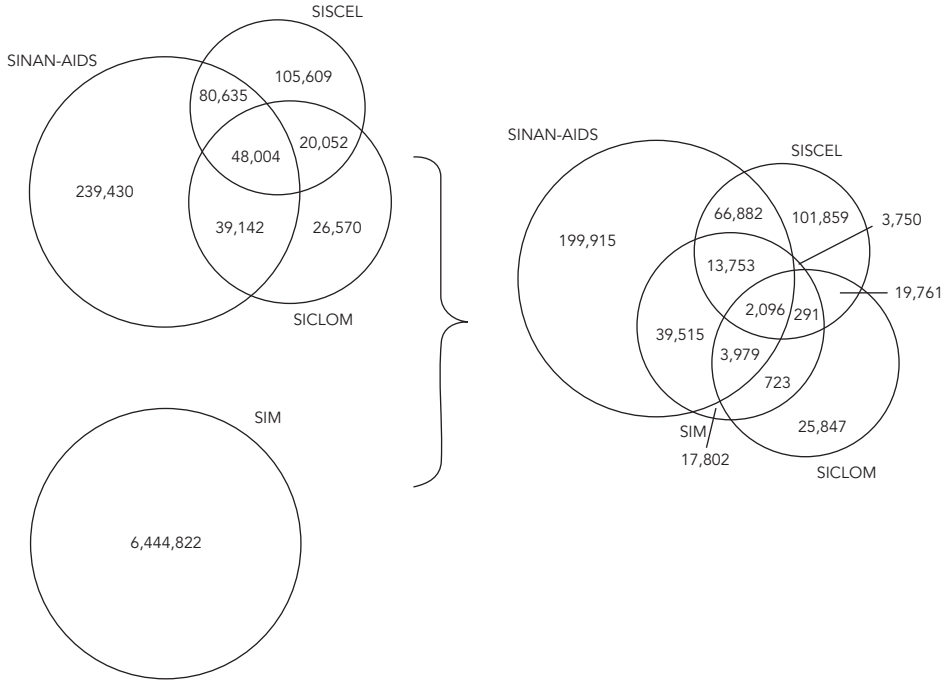
We calculated standard measures of validity (sensitivity, specificity and positive predictive value) for the entire sample. In addition, we calculated sensitivity and specificity according to some putative predictors of mortality, as follows: year of diagnosis, sex, age group, race, geographical region of residency, and exposure category. To estimate the sensitivity of the record linkage strategy, we considered as known deaths cases with a date of death registered in SINAN (N = 19,750). Specificity was estimated from AIDS cases known alive, i.e. with no date of death recorded in SINAN, and found registered in SISCEL with either lymphocytes CD4+ or viral load tests in 2006 (N = 38,675). The results of sensitivity, specificity and positive predictive value were presented along with 95% confidence intervals (95%CI) calculated using the Wilson's method [17]. Data analysis was performed with CIA software, version 2.0 (University of Southampton, Southampton, UK). The study was approved by the Ethics Committee of the Evandro Chagas Clinical Research Institute of the Oswaldo Cruz Foundation.

Results

Figure 1 presents the universe of the databases described above. From the 133,768 AIDS patients registered in SICLOM, 26.8% were also registered among the 254,300 HIV individuals registered in SISCEL. Out of 254,300 individuals registered in SISCEL, 31.6% were also found among the 477,211 AIDS cases reported in SINAN from 1980

Figure 1

Distribution of people with HIV and AIDS, according to database. Brazil, 1980-2006.



SINAN-AIDS: Information System for Notifiable Diseases of AIDS Cases (from 1980 to 2006, with 407,211 AIDS cases); SISCEL: Laboratory Test Control System (from 2002 to 2006, with 173,665 HIV+ individuals); SICLOM: System for Logistic Control of Drugs (from 2002 to 2006, with 133,768 patients in treatment); SIM: Mortality Information System (from 2000 to 2006, with 6,444,822 deaths registered with all underlying causes).
Note: people not represented in the second part (SINAN-AIDS/SICLOM: 35,163; SINAN-AIDS/SISCEL/SICLOM: 45,908).

to 2006. A total of 559,442 people with HIV and AIDS were found registered in at least one of the three systems, after excluding duplicities. A total of 64,107 people with HIV and AIDS were found among the 6,444,822 deaths registered in SIM, from 2000 to 2006. Another 17,802 people having AIDS as the main cause of death were also identified only in SIM.

Through record linkage with the SIM, 17,448 deaths of AIDS cases reported to SINAN were identified. In 17,310 cases, the death had been previously reported to SINAN, with 93% of agreement between the dates of death recorded in both databases. Thus, record linkage identified 17,310 of the 19,750 AIDS cases with date of death registered in SINAN (known death), yielding a sensitivity of 87.6% (95%CI: 87.1-88.2). Among the 38,675 AIDS cases, which were found registered in SISCEL in 2006 (known alive), record linkage erroneously classified only 138 cases as deceased (a specificity of 99.6%; 95%CI: 99.6-99.7). The positive predictive value for the entire sample was 99.2% (95%CI: 99.1-99.3). Among the 138 cases erroneously found in SIM, 2.2% and 8% had data of birth and mother's name missing, respectively, compared to 0.8% and 5%, respectively, among the 17,310 cases registered as dead in SINAN and found in SIM.

Table 1 depicts the sensitivity and specificity of the record linkage process according to year of diagnosis, sex, age group, skin color, geographical region of residency, and exposure category for both sexes. No important variation was observed in sensitivity, except for cases of less than 13 years of age (77.1%), and in less extension for female (85.5%). There were high levels of specificity for all variables analyzed.

Discussion

We found a sensitivity of 87.6% and a specificity of 99.6% of the record linkage procedure used to ascertain deaths among cases reported to the Brazilian AIDS surveillance database. The nearly perfect specificity observed in our study was to some extent expected, as we adopted a linkage strategy that sacrificed the sensitivity in order to minimize the number of false positive links. We adopted such a strategy because it has been suggested that false positive errors of the outcome classification in survival analyses, even when non-differential with regards to the exposure variable, bias both the risk difference and the risk ratio to the null [15]. On the other hand, non-differential false negative errors bias the risk difference rate but not the risk ratio [15]. Moreover, unlike false negative errors, false positive errors appeared to be dependent on the size of linked databases, increasing when larger databases are employed [16].

Our results were worse than those obtained by Pacheco et al. [6] with a deterministic linkage algorithm applied to identify deaths among HIV-infected patients of two cohort studies carried out in Rio de Janeiro, Brazil (sensitivity = 96.5% and specificity = 100%). We believe that the discrepancy between this study and our own could be due to differences in the size and the data quality of the linked databases. We used very large databases generated in all Brazilian states, which were about seven times (mortality) and seventy-nine times (HIV-AIDS surveillance) bigger than the databases used by Pacheco et al. [6]. Our HIV-AIDS database came from routine epidemiological surveillance, being prone to low accuracy and completeness. Furthermore, the use of large databases increases the chance of false positive errors and makes the clerical review process a real challenge [16].

Nakhaee et al. [18] carried out a probabilistic linkage of HIV-AIDS surveillance and mortality data in Australia. By choosing weights of match pairs that maximize sensitivity and specificity, they obtained, as the best result, a sensitivity of 82% and a specificity of 92%. Their performance was worse than ours, but they had name codes, instead of full names, available in the surveillance database. The lack of this important identifier probably decreased the discriminant power of their linkage strategy. Indeed, the number and quality of the personal identifiers available to be compared, as well as the completeness of the databases to be linked, are fundamental prerequisites for the success of a record linkage process. Applying the same technique that we used in the current study, we obtained worse results linking primary data that came from a case-control study [19], a household survey [20] and a cohort study [21] to mortality, hospital admissions and live births databases, respectively. Lack of some personal identifiers available for the linkage process (e.g.: mother's name) and problems regarding the completeness of the databases might explain the poorer performance of these previous linkage processes.

Data generated in different settings are expected to present heterogeneous accuracy and completeness. Hence, it is surprising that we did not observe an expressive difference in the sensitivity and specificity measures among the Brazilian regions. The fact that we used different blocking steps and combined the automatic linkage process with an extensive clerical review may have contributed to minimize the occurrence of misclassification errors and, consequently, the

Table 1

Total number of AIDS cases reported in the Information System for Notifiable Diseases (SINAN) with date of death or found alive in 2006, sensitivity (%), specificity (%) and respectively 95% confidence interval (95%CI), according to selected variables. Brazil, 2002-2006.

	Total		Sensitivity		Specificity	
	Death registered in SINAN	Alive in 2006	%	95%CI	%	95%CI
Total	19,750	38,675	87.6	87.2-88.1	99.6	99.6-99.7
Year of diagnosis						
2002	6,177	8,308	87.0	86.1-87.8	99.5	99.4-99.7
2003	5,361	9,770	87.8	86.9-88.7	99.6	99.4-99.7
2004	4,812	10,317	87.4	86.5-88.3	99.7	99.5-99.8
2005	3,400	10,280	88.9	87.8-89.9	99.8	99.6-99.8
Sex *						
Female	6,320	16,141	85.5	84.6-86.3	99.8	99.7-99.8
Male	13,429	22,534	88.7	88.1-89.2	99.6	99.5-99.6
Age group (years) **						
Less than 13	339	1,216	77.9	73.2-82.0	100.0	99.7-100.0
13-19	239	868	84.5	79.4-88.6	99.9	99.4-100.0
20-39	10,797	24,480	88.2	87.6-88.8	99.7	99.6-99.8
40-59	7,460	11,265	87.3	86.5-88.1	99.5	99.3-99.6
60 and +	910	845	88.0	85.8-90.0	99.4	98.6-99.7
Skin color ***						
White	9,159	18,712	89.6	89.0-90.2	99.7	99.6-99.8
Black	2,219	3,721	88.8	87.4-90.1	99.7	99.4-99.8
Brown	5,218	9,187	89.1	88.2-89.9	99.6	99.5-99.7
Region of residency						
North	871	1,785	86.9	84.5-89.0	99.9	99.6-100.0
Northeast	2,445	4,393	87.2	85.8-88.5	99.3	99.0-99.5
Southeast	10,966	23,383	87.3	86.7-87.9	99.6	99.5-99.7
South	4,164	6,763	88.7	87.7-89.7	99.9	99.7-99.9
Center West	1,304	2,351	88.2	86.3-89.8	99.9	99.6-100.0
Exposure category #						
Female						
Heterosexual	5,859	14,893	85.5	84.6-86.4	99.7	99.6-99.8
IDU	310	573	85.5	81.1-89.0	99.8	99.0-100.0
Male						
Homosexual	1,093	4,508	84.8	82.6-86.8	99.6	99.4-99.7
Bisexual	744	2,747	89.0	86.5-91.0	99.6	99.2-99.7
Heterosexual	3,719	10,387	90.7	89.7-91.6	99.5	99.4-99.6
IDU	1,556	2,130	90.2	88.6-91.5	99.8	99.5-99.9

IDU: injection drug users.

* Case excluded: 1 death, sex unknown;

** Cases excluded: 5 deaths and 1 case alive, with age unknown;

*** Cases excluded: 112 deaths and 320 cases alive with other skin color; 3,042 deaths and 6,735 cases alive with unknown skin color;

Cases excluded: 127 deaths and 146 deaths and 626 cases alive with other exposure category; 6,195 deaths and 18,277 cases alive with unknown exposure category.

differences in the results of sensitivity and specificity among the regions. It also could explain the small variation in the validity measures according to other putative predictors of mortality. The only exceptions were observed among cases less than 13 years of age, which presented a slightly worse sensitivity, and for women, with slightly lower sensitivity, although some authors consider significant only differences between proportions higher than 10% [13,22]. We did not observe any important differences in the completeness of the personal identifiers in this age range. One possible explanation for the differences observed could be the existence of some children orphaned

as a result of AIDS in this group. A study carried out in Porto Alegre, Rio Grande do Sul State, Brazil [23], found that: (a) 5% of AIDS orphans were institutionalized and 46% of them were living in substitute families (with or without any defined judicial situation); (b) HIV positivity was a significant predictor of institutionalization (orphanages and small family-type units). Therefore, with the change in the family affiliation, it is plausible to hypothesize that different personal identifiers had been reported to the surveillance and mortality databases. However, a more thorough understanding of the reasons for this discrepancy should be investigated with further analysis.

Some limitations of the current study should be mentioned. First, because we did not know the HIV-infected individuals' vital status (considered the gold standard), we only included AIDS cases reported in the surveillance database in our analysis. However, the same personal identifiers, which were used for linking such cases were also available for the HIV-infected individuals without any important differences in the completeness of these variables. Therefore, we might expect to obtain sensitivity and specificity measures similar to the ones observed for the AIDS cases, although a lower positive predictive value might be expected because of the dependence of this latter measure on the prevalence of death.

Second, we assessed the validity of the record linkage strategy against an imperfect gold standard (know vital status). Ideally, we should have compared the linked data with the vital status obtained trough an active individual follow-up strategy. This strategy is feasible in the context of epidemiological studies based on small or moderately large numbers of participants [19,24,25], however the very large number of patients included in our HIV-AIDS database would make active follow up impracticable. Moreover, it is not always possible to trace all individuals; consequently the active follow up is also prone to errors [19,24,25]. Another strategy is to manually inspect a random sample of links designated as matches and non-matches by two independent reviewers with the human judgment being considered the gold standard [26]. This procedure might be time-consuming and because of its subjective nature, it is also subject to error. Using the "known vital status" ascertained trough existing secondary databases represents a more cost-effective strategy, which has been applied in a number of studies [6,18,27]. Because we used the date of death recorded in the AIDS surveillance database to classify a patient as deceased (more detailed information), it is very unlikely that an individual reported as deceased would in fact be alive. Likewise, although possible, it is unlikely that a patient recorded in 2006 in the laboratory database would in fact be deceased. If such errors had happened, our sensitivity and specificity results would be, respectively, under and overestimated.

Nevertheless, to the best of our knowledge this is the first study conducted in a less developed country to assess the accuracy of a linkage strategy to identify deaths among cases reported to a very large national HIV-AIDS surveillance database. By combining the deceased cases recorded in the surveillance database with those identified trough the record linkage strategy, it will be possible to get a better estimate of the mortality rate in our study population. Besides, as the linkage errors were non-differential with regards to the various putative predictors of mortality and the specificity obtained was nearly perfect, we expect to obtain risk ratio estimates that are minimally biased.

In conclusion, we believe that record linkage can be a powerful tool in epidemiological and health services research. Our findings suggest that even large and heterogeneous databases can be linked with satisfactory accuracy, especially for specificity. National surveillance systems can improve epidemiological analysis by adding information reaching a high degree of completeness for substantial data through record linkages with complementary sources of data. In our study, a comparison of AIDS surveillance and mortality systems at national level indicates a high degree of completeness of the AIDS surveillance system, together with a high degree of agreement of dates of death between both systems. Using the "known vital status" ascertained trough existing secondary databases represents cost-effective strategy to evaluate record linkage accuracy.

Resumo

É importante avaliar a acurácia de relacionamento de dados, já que erros podem enviesar as medidas de ocorrência e de associação de doenças. O objetivo desse estudo é verificar a acurácia da estratégia de relacionamento probabilístico de banco de dados em identificar óbitos entre casos de AIDS notificados no Sistema de Informações de Agravos de Notificação (SINAN). O banco de dados de pessoas com HIV/AIDS (N = 559.442) foi relacionado a 6.444.822 óbitos (todas as causas) registrados no Sistema de Informações sobre Mortalidade (SIM). Para estimar as medidas de acurácia, foram selecionados todos os casos de AIDS com datas de óbito registradas no SINAN-AIDS de 2002 a 2005 (N = 19.750) e 38.675 casos sabidamente vivos em 2006. A sensibilidade foi de 87,6% (IC95%: 87,1-88,2), a especificidade de 99,6% (IC95%: 99,6-99,7) e o valor preditivo de 99,2% (IC95%: 99,1-99,3). Sensibilidade foi 12% menor para os casos com menos de 13 anos. Foram observadas pequenas variações nas medidas de validação segundo algumas variáveis preditoras de mortalidade. Conclui-se que bancos de dados grandes e heterogêneos podem ser relacionados com acurácia satisfatória.

Registro Médico Coordenado; Sistemas de Informação; Síndrome de Imunodeficiência Adquirida; Mortalidade

Contributors

M. G. P. Fonseca and C. M. Coeli conceived, designed and coordinated the study, conducted the data analysis, guided the discussion of results, and drafted the manuscript. F. F. A. Lucena conducted the record linkage, assisted the data analysis, participated in the discussion of the results, and revised the manuscript. V. G. Veloso and M. S. Carvalho conceived the study, participated in the discussion of the results, and revised the manuscript.

References

1. Fonseca MGP, Bastos FI. Twenty-five years of the AIDS epidemic in Brazil: principal epidemiological findings, 1980-2005. Cad Saúde Pública 2007; 23 Suppl 3:S333-43.
2. Lucena FFA, Fonseca MGP, Sousa AIA, Coeli CM. O relacionamento de bancos de dados na implementação da vigilância da aids. relacionamento de dados e vigilância da AIDS. Cad Saúde Colet (Rio J.) 2006; 14:305-12.
3. Rede Interagencial de Informação para a Saúde. Indicadores básicos para a saúde no Brasil: conceitos e aplicações. 2ª Ed. Brasília: Organização Pan-Americana da Saúde; 2008.
4. Centers for Disease Control and Prevention. Electronic record linkage to identify deaths among persons with AIDS: District of Columbia, 2000-2005. MMWR Morb Mortal Wkly Rep 2008; 57:631-4.
5. Deapen D, Cockburn M, Pinder R, Lu S, Wohl AR. Population-based linkage of AIDS and cancer registries: importance of linkage algorithm. Am J Prev Med 2007; 33:134-6.
6. Pacheco AG, Saraceni V, Tuboi SH, Moulton LH, Chaisson RE, Cavalcante SC, et al. 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 2008; 168:1326-32.

7. Regidor E, Sánchez E, de la Fuente L, Luquero FJ, de Mateo S, Domínguez V. Major reduction in AIDS-mortality inequalities after HAART: the importance of absolute differences in evaluating interventions. Soc Sci Med 2009; 68:419-26.
8. Serraino D, Zucchetto A, Suligoi B, Bruzzone S, Camoni L, Boros S, et al. Survival after AIDS diagnosis in Italy, 1999-2006: a population-based study. J Acquir Immune Defic Syndr 2009; 52:99-105.
9. Ministério da Saúde. Boletim Epidemiológico Aids e DST 2008, Ano V, nº. 01.
10. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques. New York: Springer; 2007.
11. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. Stat Med 1997; 16:2633-43.
12. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol 2002; 31:1246-52.
13. Drumond EF, Machado CJ. Linkage entre registros do SIHSUS e SINASC: possíveis vieses decorrentes do não-pareamento. Rev Bras Estud Popul 2007; 25:191-4.
14. Programa Nacional de DST/AIDS, Secretaria de Vigilância em Saúde, Ministério da Saúde. Critérios de definição de casos de Aids em adultos e crianças. Brasília: Ministério da Saúde; 2004. (Série Manuais, 60).
15. Camargo Jr. KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cad Saúde Pública 2000; 16:439-47.
16. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 1966; 10:707-10.
17. Altman DG, Machin D, Bryant TN, Gardner MJ, editors. Statistics with confidence: confidence intervals and statistical guidelines. 2nd Ed. London: BMJ Books; 2000.
18. Nakhaee F, McDonald A, Black D, Law M. 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 2007; 31:308-12.
19. Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida. Cad Saúde Pública 2006; 22:2249-52.
20. Coeli CM, Blais R, Costa MDCE, Almeida LM. Probabilistic linkage in household survey on hospital care usage. Rev Saúde Pública 2003; 37:91-9.
21. Coutinho RG, Coeli CM, Faerstein E, Chor D. Sensibilidade do linkage probabilístico na identificação de nascimentos informados: estudo Pró-Saúde. Rev Saúde Pública 2008; 42:1097-100.
22. Ford JB, Roberts CL, Taylor LK. Characteristics of unmatched maternal and baby records in linked birth records and hospital discharge. Paediatr Perinat Epidemiol 2006; 20:329-37.
23. Shannon HS, Jamieson E, Walsh C, Julian JA, Fair ME, Buffet A. Comparison of individual follow-up and computerized record linkage using the Canadian Mortality Data Base. Can J Public Health 1989; 80:54-7.
24. Computerized record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. The West of Scotland Coronary Prevention Study Group. J Clin Epidemiol 1995; 48:1441-52.
25. Doring M, França Junior I, Stella IM. Factors associated with institutionalization of children orphaned by AIDS in a population-based survey in Porto Alegre, Brazil. AIDS 2005; 19 Suppl 4:S59-63.
26. Qayad MG, Zhang H. Accuracy of public health data linkages. Matern Child Health J 2009; 13:531-8.
27. Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. J Am Med Inform Assoc 1997; 4:233-7.

Submitted on 27/Aug/2009
Final version resubmitted on 06/Apr/2010
Approved on 16/Apr/2010