Daniele Pinto da Silveira[I]

Elizabeth Artmann[II]

Accuracy of probabilistic record linkage applied to health databases: systematic review

ABSTRACT

OBJECTIVE: To analyze both national and international literature on validity of record linkage procedure of health databases focusing on quality assessment of results.

METHODS: A systematic review of cohort, case-control, and cross-sectional studies that evaluated quality of probabilistic record linkage of health databases was conducted. Cochrane methodology of systematic reviews was used. The following databases were widely searched: Medline, LILACS, Scopus, SciELO and Scirus. A time filter was not applied and articles were searched in the following languages: Portuguese, Spanish, French and English.

RESULTS: Summary measures of the quality of probabilistic record linkage were sensitivity, specificity, and positive predictive value. There were identified 202 studies, and after applying the inclusion criteria, a total of 33 articles were reviewed. Only six had complete data on the summary measures of interest. The main limitations were: no reviewer to evaluate titles and abstracts; and no blinding of the article's authors in the review process. Most scientific publications in this field were from the United States, United Kingdom, and New Zealand. Overall, the accuracy of probabilistic record linkage of databases ranged from 74% to 98% sensitivity and 99% to 100% specificity.

CONCLUSIONS: Probabilistic record linkage of health databases has notably been characterized by high sensitivity and greater flexibility of the procedure's sensitivity, indicating concern with data accuracy. The positive predictive value in studies shows a high proportion of truly positive record pairs. The quality assessment of these procedures has been proved essential for validating the results obtained in these studies, and can also contribute to improve large health databases available in Brazil.

DESCRIPTORS: Information Systems. Models, Statistical. Information Management. Statistical Databases. Interinstitutional Relations. Review.

INTRODUCTION

There has been an increasing number of studies aiming to the development and improvement of record linkage procedures since the 1980s, largely conducted in the United States, United Kingdom, and New Zealand.[1,5,6,12] In Brazil, despite extensive dissemination and application of these procedures in studies of different areas of knowledge, especially in epidemiology, there are still few studies aiming to identify the same individual in two or more record databases.

In recent years, the Brazilian *Ministério da Saúde* (Ministry of Health) has developed major information systems with outstanding progress toward electronic dissemination of data on births, deaths, notifiable diseases, hospital and

[I] Programa de Pós-Graduação em Saúde Pública. Escola Nacional de Saúde Pública Sergio Arouca (ENSP). Fundação Oswaldo Cruz (Fiocruz). Rio de Janeiro, RJ, Brasil

[II] Departamento de Administração e Planejamento. ENSP. Fiocruz. Rio de Janeiro, RJ, Brasil

Correspondence:
Elizabeth Artmann
Escola Nacional de Saúde Pública Sérgio Arouca
R. Leopoldo Bulhões, 1480, 7o andar – Manguinhos
21041-210 Rio de Janeiro, RJ, Brasil
E-mail: artmann@ensp.fiocruz.br

Received: 10/12/2008
Revised: 4/8/2009
Approved: 4/15/2009

outpatient care, basic health care, and public health budgets, among others.[a] However, the wide variety of system designs and lack of a univocal identifier that is able to integrate several databases make record linkage in Brazil a very complex procedure. Health information is produced and used in Brazil in a complex background of institutional relationships consisting of diverse management and financing mechanisms. Thus, integration of different databases occurs in a discontinuous and uncoordinated manner between the different levels of govern and actors involved.

Record linkage can be defined as an area of knowledge that study an approach for searching pairs or duplicated records in the same file or between files. It is usually carried out using probabilistic procedures that apply pairing of two (or more) databases using probabilities of agreement and disagreement between a set of common variables to both databases.[1] They are often used for identifying individual variables such as name, address and date of birth. Additional information such as income, education, among others, can be used depending on the information quality of these fields.[b]

The main objective of probabilistic record linkage is to find pairs of records that refer to the same individual as well as to standardize data and assess information quality.[10,c] But low quality of constant data in health information systems can make it difficult record linkage procedures or lead to errors in pairing of variables. Hence, it is essential to assess the accuracy of methods used for linkage of medical records, vital statistics and large databases to increase ability to identify individual records and improve information quality and reliability. One approach to assess accuracy of record linkage procedures is to perform studies of accuracy.

In epidemiology, accuracy is considered a measure of validity that is quite often applied in studies assessing diagnostic and screening tests.[7,14] Studies of accuracy allow to assessing to which extent data is actually measuring what it is supposed to be measured or to what extent the results of a measurement are consistent with the true status of the phenomenon measured. Accuracy of a test or a method is commonly based on a gold-standard. However, gold-standards are not always available for comparison, either in terms of a new medical test or a new method to be applied. Sometimes it is required to choose as a standard of validity another method that is known to be flawed.[7] Thus, to assess performance of record linkage procedures, the results obtained in the linkage process have to be compared to an independent source of information on the occurrence of outcomes of interest (gold-standard). However, these sources are very limited.[8] When there is no gold-standard to determine specificity and sensitivity of a record linkage, the quality of linkage can be assessed only through indirect measures. Some authors have reported the use of these measures.[1,16] Blakely et al (1999)[d] used the proportion of records in a mortality database linked to records of another database at each step of the linkage procedure to estimate the number of false-positive using the number of duplicated links in both databases.

Sensitivity and specificity are classical measures of validity used when exposure and outcome are categorical variables. In epidemiology, sensitivity refers to the proportion of individuals who have an outcome of interest that are classified as positive in a test while specificity refers to the proportion of individuals that do not have a disease or outcome of interest and are identified as negative in the test. In the studies of accuracy of record linkage, accuracy can also be assessed in terms of the method's sensitivity, specificity, and positive predictive value, as per analogy, the true pair of linkage (match) can be considered equivalent to the presence of the outcome of interest in epidemiological studies (e.g., death).[1,14]

The objective of the present study was to review the Brazilian and international literature on validity of record linkage procedures applied to health databases, focusing on measures used to assess quality of results.

METHODOLOGICAL PROCEDURES

A retrospective systematic review of cohort, case-control, and cross-sectional studies was carried out with the main objective of assessing accuracy of probabilistic record linkage in health databases. Cochrane methodology of systematic reviews was used.[9] Articles were identified through literature search in Medline, LILACS, Scirus, SciELO and Scopus databases between November and December 2007. A time filter by year of publication was not applied due to the fact that we assumed that there would be few publications and major articles could be missed. The single limiting factor was Medline's filter that only searches articles from 1966.

The selected strategies and key words for searching the databases were: (*record linkage*) OR (*record linkage*

[a] Pan American Health Organization. Rede Interagencial de Informações para Saúde (Ripsa). Indicadores básicos para a saúde no Brasil: conceitos e aplicações. Brasília; 2002.

[b] Winkler WE. Automatically Estimating Record Linkage False Match Rates. Washington: Statistical Research Division United States Census Bureau; 2007.

[c] Shah GH, Taima F, McBrida S. Probabilistic linkage in Public Health: Results of the NAHDO Survey. A critical assessment of record linkage software used in public health. Salt Lake City: National Association of Health Data Organizations; 2008[citado 2008 jan 21]. Disponível em: http://www.nahdo.org/.

[d] Blakely T, Salmond C, Woodward A. Anonymous record linkage of 1991 census records and 1991-94 mortality records: The New Zealand Censu-Mortality Study. Wellington: Department of Public Health, School of Medicine, University of Otago; 1999.

AND *health*) OR (*record linkage* AND *accuracy*) OR (*record linkage* AND *health information*) OR (*record linkage* AND *information system*) OR (*record linkage* AND *accuracy* AND *probabilistic* AND *specificity* AND *health data*) OR (*record linkage* AND *health* AND *accuracy*) OR (*medical record linkage*) OR (*medical record linkage* AND *accuracy*) OR (*medical record linkage* AND *health*). There were selected articles published in Portuguese, Spanish, English, and French language.

Then a second strategy was used to manually search lists of reference of the articles that were identified and selected.

The studies had as outcome measures of accuracy of probabilistic record linkage: sensitivity, specificity, and positive predictive value. As in epidemiological studies that usually use these measures to assess accuracy of medical tests, many authors have been applying these concepts to record linkage.[1,5,6,12,16] Sensitivity in studies of accuracy was defined as the proportion of all records in a file or database that have a match in another file correctly accepted as links (true-positive). Specificity was the proportion of all records in a file that *do not* have matches in another base correctly not accepted as links (true-negative). The positive predictive value was measured in studies of accuracy by means of the occurrence of duplicated links (e.g., a record of death linked to two or more records of another comparison database). This occurrence can be used to estimate the positive predictive value of the result of classification criteria (decision model).

In the international literature, match is defined as a pair of records belonging to the same individual while link is a pair that is accepted as a probable "true" pair in the record blockage step when small blocks of records are created and compared between them.[1]

No reviewer evaluated the titles and abstracts of all studies found in the electronic search and the articles' authors were not blinded in the review process.

Each study was checked for the main inclusion criteria: type of design, sample universe (or sample size), database size, sensitivity, specificity, and positive predictive value. A standardized form was used to collect information from the articles.

There were excluded from the analysis articles using deterministic record linkage.

A convenience sample was used following the inclusion and exclusion criteria defined in the study.

ANALYSIS AND DISCUSSION OF RESULTS

In the literature search there were found 180 articles, of which 47 were selected based on title and abstract reading. Only 33 articles were selected for review, and 14 were excluded as they did not meet the previously defined criteria. Four studies were identified in the references of the articles. All studies were categorized by study design. Cross-sectional studies were 54% (n=18) of all articles reviewed while cohort studies were 24% (eight). The majority of them were retrospective cohort studies (n=7).

Most studies had high sensitivity around 93% to 99%, although there were also studies with lower sensitivities (74% to 83%).

After fully reading all articles, few studies had data on specificity (n=5) and positive predictive value (n=4) related to the linkage procedure. Most articles (n=28) only had information on method sensitivity possibly because other measures were not easily assessed. All measures were made in most studies when: a) another correlated database was available to validate data identified using a probabilistic procedure; b) epidemiological results were available (such as in cohort and cross-sectional studies) for comparison and estimate false-negatives (pairs that were true but were classified as false); and c) parameters were estimated that allowed to inferring the method specificity (application of Bayesian statistical methods). Given these limitations, it was opted to present only detailed data of articles with complete information on sensitivity, specificity, and positive predictive value (Table).

The articles were grouped by period of years to facilitate an analysis of the number of studies over the years. A larger number of studies were published from the year 2000. With respect to the sample universe, most studies (n=31) used hospital or population-based databases in addition to health-related administrative records (n=6) (Figure).

More publications on the assessment of probabilistic record linkage were from: United States (n=8), New Zealand (n=6), and United Kingdom (n=6). There were four studies from Brazil. All remaining studies were from European countries.

In regard to database size, most studies were based on middle-size databases ranging between 100,000 and 700,000 records. Many studies also reported using small vital statistics databases with a sample size of three to four digits.

The study with the largest database was Victor & Mera,[15] conducted in the United States, which linked 1.7 million administrative health insurance records to 8.5 million health care records, reporting 92% sensitivity in the linkage procedure.

In Brazil, the most widely used program in these studies was RecLink.[3] In other countries, various different record linkage programs were used such as Automach® and Statistics Canada's Generalized Record Linkage

Table. Characteristics of the studies analyzed on accuracy of probabilistic record linkage applied to health databases from 1999 to 2006.

Author/year	Country	Study design	Database size	Sensitivity	Specificity	Positive predictive value
Ellekjaer et al, 1999	Norway and Sweden	Cross-sectional	70,000 population-based records of cerebrovascular disease and 759 hospital discharge records	86%	Not available	68%
Blakely & Salmond, 2002	New Zealand	Cross-sectional	3,131,176 records from New Zealand Census-Mortality Study and 39,515 mortality records from 1986 to 1989	Not available	Not available	93% to 99%
Grannis et al, 2003	US	Cross-sectional	2 pairs of files including 6,000 Social Security Death Master File records	99.2% (1st record) 99.0% (2nd record)	99.4% (1st record) 99.4% (2nd record)	Not available
Zingmond et al, 2004	United Kingdom	Retrospective cohort	1,858,458 hospital discharge records from California (US) and 69,757 hospital death records (1990 to 1999)	95%	99%	99%
Coutinho & Coeli, 2006	Brazil	Prospective cohort	250 hospital records of a cohort of elderly patients admitted due to fracture in the city of Rio de Janeiro and death records in the State of Rio de Janeiro (n of death records not available)	86%	99%	98%
Nagle et al, 2006	New Zealand	Retrospective cohort	822 records of women diagnosed with ovary cancer between 1990 and 1993 from the National Death Index (NDI/Australia), population-based death records and 450 deaths from NDI	93%	100%	Not available

System, developed by the Canadian government, in addition to other computer programs.

Among studies assessing method specificity, it ranged between 99% and 100% while positive predictive values ranged between 68% and 99%.

Blakely & Salmond[1] and Zigmond et al[17] studies showed the highest positive predictive values. Using a computer program, Zigmond et al linked 1,858,458 hospital discharge records with 69,757 hospital death records, and obtained high sensitivity (95%) and high specificity (99%) of the linkage method as well. In Brazil, Coutinho & Coeli,[5] using quite smaller databases (n=250 hospital records), found similar results: 99% specificity and 98.10% positive predictive value.

FINAL CONSIDERATIONS

Probabilistic record linkage has been widely applied in public health in the last 50 years since the publication of Newcombe et al study.[11] In recent decades, the assessment of results obtained using record linkage has gained importance and several studies have been conducted to estimate accuracy of these procedures.

The procedures and methods of record linkage do not always provide all the required information for researchers to decide whether a pair of records is actually a true pair (match). Sometimes they need additional information besides that provided by pairing variables. Researchers most commonly recur to manual review of records, which is regarded as the gold-standard in the international literature. But as large databases are often used in health studies, manual review is not a feasible option because it makes the validation of record linkage results a costly, time-consuming procedure.[10]

Studies on accuracy of record linkage procedures faces the challenge of finding a reference database that can be used as a parameter of comparison and confirmation of status of a given record in another database. In addition to this difficulty, database quality also affects the linkage process. It is an additional element that should be taken into consideration while planning studies of accuracy because the fields used in the process of pairing variables may not appropriately filled out or validated.

In many studies, probabilistic record linkage accuracy is strongly dependent on the number and quality of fields available for comparison. Few fields available can increase the occurrence of false-positive pairs, and the classification of pairs as true when they refer to different individuals in the compared databases.[6]

Overall, it was noted in the present study that database size was not necessarily associated to the method sensitivity result. This finding seems to corroborate the assumption that the method accuracy is more directly associated to the quality of records and fields that are used in probabilistic record linkage.

Also, some authors such as Brenner et al[2] have pointed that homonymous errors tend to increase as the number of records in the databases used for linkage grows.

Camargo & Coeli[3] have noted that lower positive predictive value in record linkage of large databases can be associated to factors such as reduced prevalence of true pairs.

Some authors have claimed that the positive predictive value is the most adequate measure to assess quality of a record linkage procedure.[2,4] But, as it is virtually impossible to review all record pairs, the solution would be to generate a sample and review only some pairs. However, Sauleau et al[13] have argued that, although this is a feasible solution, it does not fully serve the purpose of assessing quality of data generation process resulting from record linkage.

The most adequate indicator to assess accuracy of probabilistic record linkage would be the percent of duplicated records.[13] Blakely & Salmond[1] have also pointed that the occurrence of duplicated links in the linkage process can be used to measure the positive predictive value of the method.

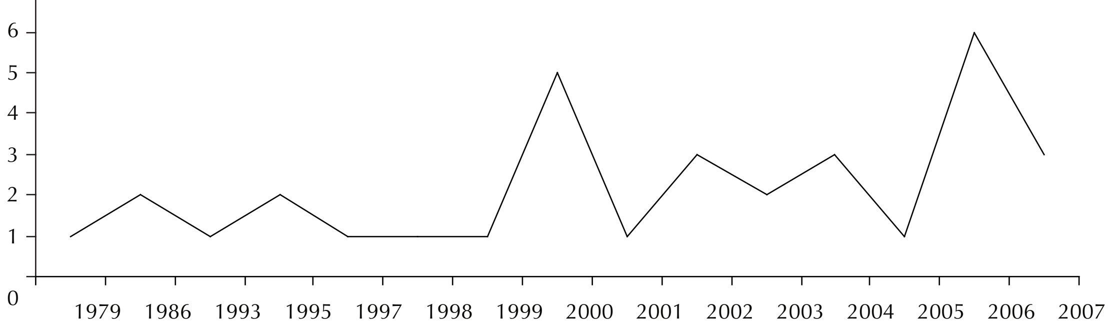

Figure. Time series of scientific publications on accuracy of probabilistic record linkage of databases from 1979 to 2008.

Other authors have claimed that the sensitivity of linkage procedure can also be estimated by adopting as gold-standard total true pairs identified in both automated search using specific programs and manual search. In some cases, confidence intervals can be added to the estimate.[3]

Camargo & Coeli[3] showed in their study that sensitivities of manual and automated processes were very similar when databases with smaller number of records were used. But as databases increased in size, there was a reduction in manual review but not in automated record linkage.

Therefore, when there are no gold-standards to establish sensitivity and specificity of record linkage, the quality of linkage can be assessed only through indirect measures. Some of these measures were used in a study conducted by Blakely et al,[a] for example, percent of records in a database identified in another database at each step of the linkage procedure and percent of records in a database identified in a second database in the final linkage process.

For these authors,[a] the total percent is close to the sensitivity of probabilistic record linkage, that would be equal to the number of true links (matches) identified, divided by the total number of true pairs. Specificity would be the number of incorrect links rejected, divided by the total number of incorrect links.

The method described by Blakely & Salmond[1] to estimate the number of false-positive (Duplicate Method) is applicable only when there is a true pair (match) for a given record, e.g., in epidemiological studies, for example, in the linkage of mortality records with other databases. This method measures the number of false-positives above a given total weight using the number of duplicated links found above this score. The occurrence of duplicated links can be used to assess the positive predictive value. This measurement supports an informed decision on the weight of the cutoff value above which the links will be accepted.

Grannis et al[8] support that the probabilistic record procedure is an advance compared to the deterministic method for several reasons. It yields an increased sensitivity by around 7% with a small reduction in specificity. The sensitivity of deterministic methods, although it can be 100%, it can be significantly reduced when data with different identification characteristics, such as different ethnic name, are used.

In the light of all information presented, it is stressed the relevance of using and improving probabilistic record linkage procedures in collective health. The quality assessment of the methods used has proven crucial to validate the results of these studies, and can also contribute to the qualification of large health databases available in Brazil. Further studies on the accuracy of linkage procedures in epidemiological studies are needed in Brazil.

[a] Blakely T, Salmond C, Woodward A. Anonymous record linkage of 1991 census records and 1991-94 mortality records: The New Zealand Censu-Mortality Study. Wellington: Department of Public Health, School of Medicine, University of Otago; 1999.

REFERENCES

1. Blakely T, Salmond C. Probabilistic Record Linkage and a method to calculate the positive predictive value. *Int J Epidemiol.* 2002,31(6):1246-52. DOI: 10.1093/ije/31.6.1246

2. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. *Stat Med.* 1997;16(23):2633-43. DOI: 10.1002/(SICI)1097-0258(19971215)16:23<2633::AID-SIM702>3.0.CO;2-1

3. Camargo Jr KR, Coeli CM. RecLink: aplicativo para o relacionamento de base de dados, implementando o método probabilistic record linkage. *Cad Saude Publica.* 2000;16(2):439-47. DOI: 10.1590/S0102-311X2000000200014

4. Coeli CM, Blais R, Costa MCE, Almeida LM. Probabilistic linkage in household survey on hospital care usage. *Rev Saude Publica.* 2003;37(1):91-9. DOI: 10.1590/S0034-89102003000100014

5. Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida. *Cad Saude Publica.* 2006;22(10):2249-52. DOI: 10.1590/S0102-311X2006001000031

6. Coutinho RGM, Coeli CM, Faerstein E, Chor D. Sensibilidade do linkage probabilístico na identificação de nascimentos informados: Estudo Pró-Saúde. *Rev Saude Publica.* 2008;42(6):1097-100. DOI: 10.1590/S0034-89102008005000053

7. Fletcher R, Fletcher S. Epidemiologia clínica: Elementos essenciais. 4.ed. Porto Alegre: Artes Médicas; 2004.

8. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. *AMIA Annu Symp Proc.* 2003:259-63.

9. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions 4.2.6. Chichester: John Wiley & Sons; 2006[citado em 2007 ago 04]. (The Cochrane Library, 4). Disponível em: http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.pdf

10. Machado CJ, Hill K. Probabilistic Record Linkage and an automated procedure to minimize the undecided-matched pair problem. *Cad Saude Publica.* 2004;20(4):915-25. DOI: 10.1590/S0102-311X2004000400005

11. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science.* 1959;130:954-9. DOI: 10.1126/science.130.3381.954

12. Roos LL, Wajda A. Record linkage strategies. Part I: estimating information and evaluation approaches. *Methods Inf Med.* 1991;30(2):117-23.

13. Sauleau EA, Paumier JP, Buemi A. Medical Record Linkage in health information systems by approximate string matching and clustering. *BMC Med Inform Decis Mak.* 2005;5:32. DOI: 10.1186/1472-6947-5-32

14. Sklo M, Nieto FJ. Epidemiology: Beyond the Basics. London: Jones and Bartlett Publishers; 2004.

15. Victor TW, Mera RM. Record linkage of health care insurance claims. *J Am Med Inform Assoc.* 2001;8(3):281-8.

16. Winkler WE. Record linkage: overview of recent developments and applications. In: Falorsi P, Pallara A, Russo A, editors. L'integrazione di dati di fonti diverse, Technice e applicaczioni del Record Linkage e metodi di stima basati sull'uso congiunto di fonti statistche e amministrative. Rome: Franco Angeli Editore; 2005.

17. Zingmond DS, Ye Z, Ettner SL, Liu H. Linking hospital discharge and death records accuracy and sources of bias. *J Clin Epidemiol.* 2004;57(1):21-9. DOI: 10.1016/S0895-4356(03)00250-6

Study presented at the V Annual Meeting HTAI 2008 – Health Technology Assessment in Context, Montreal, Canada, 2008.