Rev Saúde Pública 2011;45(2)

Arn Migowski[I]

Rogério Brant Martins Chaves[I]

Cláudia Medina Coeli[II]

Antonio Luiz Pinho Ribeiro[III]

Bernardo Rangel Tura[I]

Maria Cristina Caetano Kuschnir[I]

Vitor Manuel Pereira Azevedo[I]

Daniel Brasil Floriano[I]

Carlos Alberto Moreira Magalhães[I]

Márcia Cristina Chagas Macedo Pinheiro[I]

Regina Maria de Aquino Xavier[I]

Accuracy of probabilistic record linkage in the assessment of high-complexity cardiology procedures

ABSTRACT

OBJECTIVE: To evaluate the viability of a probabilistic record linkage strategy to identify deaths of patients who underwent complex cardiology procedures among the total deceased population.

METHODS: The processing cost was estimated based on 1,672 records of patients undergoing coronary artery bypass grafting that were compared with all death records in Brazil in 2005. The accuracy of the relationship was based on the probabilistic linkage of 99 hospital admissions records of patients, with known vital status, who underwent cardiac surgery at a single cardiology institute, with the death records of the state of Rio de Janeiro, Southeastern Brazil, in 2005. Linkage was conducted in four stages: standardizing the databases, blocking, matching, and rating peers. Blocking in five steps was used, with blocking keys formed by a combination of variables such as soundex codes for the first and last names, sex, and year of birth. The variables used for pairing were "full name" with the use of Levenshtein distance and "birth date".

RESULTS: The second and fifth blocking steps resulted in the largest number of formed pairs and the largest processing times for the matching. The fourth step required a lower processing cost. In the accuracy study, after five blocking steps, the sensitivity of the linkage was 90.6%, and the specificity was 100%.

CONCLUSIONS: The probabilistic strategy used has high accuracy and can be used in studies of the effectiveness of high-complexity, high-cost cardiology procedures.

DESCRIPTORS: Cardiovascular Surgical Procedures. Heart disease, mortality. Records as Topic. Mortality registries. Information Systems. Medical Record Linkage.

INTRODUCTION

Due to the high costs, the incorporation of new technologies such that the assessment of technologies used on a large-scale is an area of great interest for the Healthcare Systems worldwide. The use of technologies outside of conditions in which they are proven effective is another field of growing interest.[14]

The incrementally growing costs associated with high-complexity cardiovascular procedures caused the total cost of the Sistema Único de Saúde (SUS – Unified Health System) to increase from about 395 million Reais in 2000 to around 736 million Reais in 2005, an increase of 86.2%, according to DATASUS (Information Department of SUS). Almost 52 million Reais were spent on the

[I] Núcleo de Saúde Coletiva. Coordenação de Ensino e Pesquisa. Instituto Nacional de Cardiologia. Rio de Janeiro, Brasil

[II] Instituto de Estudos em Saúde Coletiva. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

[III] Departamento de Clínica Médica. Faculdade de Medicina. Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brasil

Correspondence:
Arn Migowski
Núcleo de Saúde Coletiva
R. das Laranjeiras, 374, 5º andar
Laranjeiras
22240-006 Rio de Janeiro, RJ, Brasil
E-mail: arnmigowski@yahoo.com.br

Received: 12/17/2009
Approved: 8/25/2010

Article available from: www.scielo.br/rsp

treatment of 1,549 patients who were candidates for a high-cost pacemaker implant (implantable defibrillator and/or multisite pacemaker) in 2005. However, the medium- and long-term effectiveness of such specialized cardiology interventions promoted and financed by the SUS are not known.

The Sistema de Informações Hospitalares (SIH-SUS – Hospital Information System) is the main source of information on high-complexity cardiovascular procedures paid for by SUS. SIH-SUS is used to study factors related to access to and the quality of highly complex cardiology care.[a,b]

Despite the limitations of this database, which is primarily administrative, the reliability of data from hospital admission authorizations (HAAs) on the diagnosis of cardiovascular diseases is considered satisfactory.[7,10]

Limitations of the SIH with respect to information on the date and cause of death has been overcome using data from the Sistema de Informações de Mortalidade (SIM – Death Information System) through the identification of deaths after hospital discharge and in longer follow-up periods or due to the improvement on the underlying cause-of-death data.[11]

Due to the lack of a unique identifier field capable of uniting the records of the SIH and the SIM in a deterministic way, we cannot take advantage of the complementarity of the two systems. To minimize such problems, procedures to relate entries in several databases have been developed, many of which use probabilistic methods.[2,13]

However, few studies have assessed the accuracy of probabilistic record linkage strategies when using national databases, and no study has been conducted to assess the linkage between the SIM and the SIH-SUS in the context of complex cardiovascular procedures.[15] The processing cost of the probabilistic algorithm, regardless of its impact on the feasibility of the method, has also been rarely studied.

This study evaluated the feasibility of a probabilistic record linkage strategy to identify deaths of patients who underwent complex cardiology procedures among the total deceased population.

METHODS

The data sources used were HAAs of the SIH-SUS from 2005, as the reference database, and the database of death certificates from the SIM for the same year, as the comparison database.

Two studies were conducted. In one, we assessed the processing cost and adequacy of the blocking strategy. The other study assessed the accuracy of the probabilistic record linkage.

To evaluate the processing cost and efficacy in identifying true pairs at each step of blocking, we selected 1,672 patients undergoing coronary artery bypass grafting (CABG) with HAA paid by SUS in Brazil in 2005 and who died during hospitalization.

CABG was chosen because it is a relatively frequent procedure. Only in-hospital deaths were included in this study because there was no gold standard to determine the vital *status* of the group of patients after discharge. Isolated CABGs were considered in addition to CABGs that were performed in combination with other procedures, such as valve replacement or infarctectomy and aneurysmectomy. The file of death certificates contained 1,006,827 records, representing all deaths in Brazil in 2005.

The number of possible pairs with the combination of these two databases would be equivalent to the product of the number of records from the two bases (1,683,414,744 pairs). With blocking, comparisons are limited to records from the same block, or, in other words, to those records that have the equal values for all variables contained in each blocking key, which reduces the total number of pairs formed at each step and increases the probability of the formation of true pairs.[5] Blocking reduces the number of comparisons made by computer and memory and processor usage, which results in lower processing costs and makes the process faster and less demanding with respect to the computing hardware. A very large number of pairs could make clerical review too complex to warrant doing.

To evaluate the accuracy of the probabilistic strategy, 452 consecutive patients were selected from the cardiac surgery database of the cardiology institute in 2005. We performed a deterministic linkage with the HAA data presented by the institute for the selected period, and 353 records were located. For 128 of these records, we could not identify the number of the HAA corresponding to the procedure of interest because these individuals had more than one HAA; excluding these individuals left 225 patients. Telephone calls were conducted to ascertain the vital *status* of these 225 patients in March and April of 2008 or the date

[a] Noronha JC. Use of result indicators for quality assessment in acute hospitals: hospital mortality after coronary artery bypass grafting in Brazilian hospitals [PhD thesis]. Rio de Janeiro: Instituto de Medicina Social, UERJ, 2001.
[b] TM Lyra. The challenge of equity in the public healthcare system: the use of hospital information systems to assess the distribution of high complexity cardiac care, Brazil 1993-1999 [Master's dissertation]. Recife: Aggeu Magalhães Research Center Fiocruz, 2001.
[c] MySQL The world's most popular open source database [Internet]. Sweden: Oracle, 2009 [cited 2009 Aug 6]. Available from: http://www.mysql. com
[d] Delphi 7: Desktop Tool Database [CD-ROM]. Version 7.0. Austin: Borland International Inc, 1992.

and place of death in cases in which the patient had already died.

Of the total, only 102 (45%) were located in an active search, despite several attempts. Of these, eight had died in the intervening period (two at home and six in the hospital), and 94 were alive. A deterministic linkage was conducted to locate the HAAs. The number of HAAs was used to pair data from the active search to those of the SIH in 2005, which was provided by DATASUS.

We located 45 of the 102 patients (44%) in the DATASUS database: 44 were alive in 2008, and one had died during hospitalization. To these 44 patients, we added 55 patients who underwent cardiac surgery with known hospital death as recorded in the HAA files, yielding 99 patients.

The HAA files were provided by DATASUS by month and federal unit, and a new database was created in MySQL.[c] The selection of variables from the SIM and SIH-SUS files and the sets of records that were used in each step of the linkage process, the homogenization of size, and the codification of the variables of the two databases were performed using Delphi software.[d]

The probabilistic linkage was carried out employing the third version of the software RecLink.[2,3] Before the linkage, the databases were standardized to achieve uniform spelling, format, and content. A five-step blocking strategy was used based on the combination of the following fields: first name Soundex code (modified); last name Soundex code (modified), sex, and year of birth.[5]

Next, we performed the matching, in which pairs were formed (an SIH record with a SIM record) from the comparison of previously selected variables. To this end, we used the variables "full name" and "birth date". Scores were generated for each pair based on the similarity of the values of the selected variables. For comparison of the variable "full name" between pairs, the Levenshtein distance was used, and for the variable "birth date", we used an algorithm for character.[2] The parameters for the construction of weighting factors (agreement or disagreement) were estimated based on the expectation-maximization algorithm in RecLink.[12]

Two researchers manually reviewed and classified the pairs as "true", "doubtful", or "false". An extensive review of all pairs was performed for pairs with higher scores, for pairs with identical names and intermediary scores, and for pairs with identical dates of birth and low scores. Pairs with very low scores were automatically considered false.

Initially, a set of preliminary criteria for the classification of pairs as true or false was established. The criteria used involved the rarity of the names and surnames and the numbers of matching names, dates of birth and county of residence.

The records of the true pairs were drawn from two databases (Table 1), and the second blocking step was performed, followed by new matching and new clerical review. The following steps were carried on until the fifth blocking step.

The computer used had the following features: quad-core processor with 45 nm technology a clock speed of 2.83 GHz, 8 GB of memory, 1333 MHz Dual Channel and a 300 GB SATA II hard drive with a speed of 10,000 rpm. Using a software performance test trial of PassMark Software Inc., this computer obtained a score of 1,306. This score indicates that the computer performs its operations 1,306 times faster than the ATX 286, the computer used as a basis for comparison, and enables the benchmarking with other computers.

The indicators used to assess the processing cost were the number of pairs found in each step and the time spent in each stage of the process. In addition to reflecting the cost of processing, the number of pairs found in each step is an indicator of the efficiency of the blocking stage.

We classified the total number of pairs formed in each step as "true pairs", "gray zone pairs", or "false pairs". The "gray zone" refers to bands of scores that contained scores that could not be immediately classified as true or by virtue of its internal heterogeneity. As a result, clerical review was required for each pair in these bands. Therefore, the number of pairs in the gray zone (Table 2) was an indicator of the manual workload required to classify the questionable pairs.

To evaluate the accuracy of the method, we used sensitivity and specificity and positive and negative predictive values. The information on vital *status* obtained for each patient in the active search or from documented hospital mortality in the HAAs were considered the gold standard. A gold standard is defined as an external source of "truth" about the value of the variable of interest - in this case the vital *status* - for each individual in the study population.[8] In the clerical review, the researchers were blinded to the vital *status* given by the gold standard.

Two HAAs with nonspecific names beginning with "RN of" (newborn of) followed by the name of the mother were excluded from the accuracy study because of the inability to locate these patients in the mortality database. Ninety-five percent confidence intervals were calculated by the Wilson[1] method with the PropCIs package version 0.1-6 for the software R.

This research was previously approved by the Ethics Committee of the Universidade Federal de Minas Gerais (on 5/4/2009, Protocol 0084.0.203.000-09).

RESULTS

The second and fifth steps were those with the highest execution time of matching because they had less specific blocking keys (Table 1). This result was confirmed by the greater number of pairs found in these two steps (Table 2). The first and second steps, which identified the largest numbers of true pairs, required more processing time for the formation of files from the linkage (Table 1).

The selection of cut-off points and the automatic classification of pairs, called "step 0", took 42 minutes after the first blocking step, while the alternate strategy of clerical review of the pairs took two hours to complete (Table 1). The clerical review strategy had a sensitivity of 80%, while the "automatic" classification strategy had a sensitivity of 72% (Table 2).

In the clerical review, the location of the true pairs was not homogeneous among the dozens of bands of scores. We identified three bands of interest: (A) the first score band, in which high scores resulted from a high degree of similarity between the pairing variables (full name and date of birth); (B) a band with intermediate scores for pairs with identical names and different dates of birth; and (C) an area with low scores for pairs with identical birth dates and different names (by the Levenshtein distance). In the last type, the difference between full names could have resulted from one file containing an abbreviation or omission of intermediate names, which has a large impact on the Levenshtein distance but is easily identifiable in the clerical review.

Of the 1,411 true pairs found in the study of processing cost (Table 2), 97.9% were located in bands with the characteristics described in A, 2% in bands described in C, and 0.1% in bands of the type described in B.

In both studies, the first blocking step resulted in the greatest number of true pairs (95% of all true pairs in the study of processing cost and 96% in the study of accuracy) (Table 3).

In the study of accuracy, the sensitivity was 90.6% (Table 4). Five patients, who we know died during the study period, were not found by probabilistic record linkage. In the accuracy study, no patient known to be living at the end of the follow-up was classified as dead (false positives), resulting in a specificity of 100% (Table 4).

DISCUSSION

The present study showed good results for the accuracy of the probabilistic record linkage strategy used and had an acceptable processing cost. The number of deaths was slightly underestimated with the use of this method (five false negatives), and no patient still living at the end of follow up was classified as dead (no false

Table 1. Processing time for the linkage between the databases of information systems, according to the blocking step.

Steps	Processing Time (min)					
	Step 0[a]	Step 1	Step 2	Step 3	Step 4	Step 5
Estimation of pairing parameters	12	12	0	0	0	0
Implementation of the pairing (formation of pairs with scores)	2	2	16	8	0.27	9
Manual selection of cutoff points w / automatic classification of pairs	30	0	0	0	0	0
Automatic classification of pairs	12	0	0	0	0	0
Manual sorting of the pairs	0	120	105	20	25	15
File creation of true pairs and new files SIH and SIM	30	32	6	4	3	5
Total	88	166	127	32	28	29

[a] Step 0 is only an alternative approach to Step 1
SIH: Sistema de Informações Hospitalares (Hospital Information System)
SIM: Sistema de Informações de Mortalidade (Death Information System)

Table 2. Results of applying the blocking strategy in different steps.

Steps	Number of pairs found in the step	Number of pairs in the gray zone	Number of true pairs in the step	Sensitivity of linkage (%)
0 (same as step 1, but without manual inspection of the gray zone)	38,712	14,043	1,205	72
1: first name soundex (modified) + soundex of last name (modified) + sex	38,712	14,043	1,340	80
2: the first name soundex (modified) + sex	146,070	3,812	32	82
3: soundex of last name (modified) + sex	91,324	4,453	11	83
4: the first name soundex (modified) + soundex of last name (modified)	9,024	3,570	26	84
5: year of birth + sex	695,253	69	2	84

Table 3. Results of applying the blocking strategy in different steps in the study of accuracy.

Steps	True number of pairs in Step	Sensitivity of linkage (%)	Specificity of linkage (%)
1: first name soundex (modified) + soundex of last name (modified) + sex	46	86.8	100
2: the first name soundex (modified) + sex	1	88.7	100
3: soundex of last name (modified) + sex	0	88.7	100
4: the first name soundex (modified) + soundex of last name (modified)	1	90.6	100
5: year of birth + sex	0	90.6	100

positives), even in the absence of variables that would increase the specificity of the clerical review, such as "mother's name" and "city of birth". As the event "death" is relatively rare among patients undergoing the studied procedures, small errors in specificity would have a great impact on the quality of linkage.

The low processing time is probably associated to the efficiency of the multiple-step blocking strategy, the use of blocking keys with multiple variables, and the appropriate configuration of the computer. The number of procedures used in the study of processing cost is applicable for annual assessment of the most highly complex cardiovascular procedures except coronary angioplasty and coronary artery bypass grafting, which are more frequent.

The fourth step required considerably less processing time than the other steps did and yielded important results because this step allowed the identification of true pairs in which there was a typo in the sex variable, as this was the only blocking key without this variable.

The fifth blocking step did not substantially increase the method's sensitivity and enabled the location of only two additional true pairs, which, despite having appeared in the "gray zone" of other steps, had not been identified. Aside from being one of the more demanding processing stages because of the large number of pairs formed, the ease of identifying false pairs (e.g., records with missing data) meant that the manual inspection was not difficult. The fifth step can be used when the databases used are not very large and when the sensitivity of the linkage needs to be increased.

The clerical review was the step that consumed the most time. This stage had a great impact on the sensitivity of the method. In addition to being responsible for an 8% increase in sensitivity in the first step, it was the only method used for classifying peers in steps 2, 3, 4 and 5, depending on the impossibility of establishing cut-off points for true pairs with strong powers of discrimination in the referenced steps. The inclusion of large numbers of in-hospital deaths (55 patients) could increase the proportion of deaths found with linkage due to the possibility of a better quality of information on the death certificate, which would increase sensitivity.

Table 4. Accuracy in identifying deaths due to the probabilistic.

Passive Tracking (Linkage)	Active Tracking (gold standard)	
	Death	Living
Death	48	0
Living	5	44

Sensitivity = 90.6% (95%CI: 79.7;95.9)
Specificity = 100% (95%CI: 92.0;100)
Positive predictive value = 100% (95%CI: 92.6;100)
Negative predictive value = 89.8% (95%CI: 78.2;95.6)

The need for clerical review in linkages involving larger numbers of records or steps could derail the process. According to our results, one option would be to conduct selective inspection of the score ranges of the greatest interest, with features that are easy to identify, and which account for 100% of the pairs found. Another acceptable strategy when the sensitivity of probabilistic linkage is not critical would be to conduct only the first blocking step, given that this step identified 95% of the true pairs found in our studies. The dominant role of the first blocking step was also observed by other researchers.[5] Researchers have invested in automated techniques to reduce the number of pairs requiring clerical review.[9]

The sensitivity and specificity results of the method were similar to those of other studied methods that involved probabilistic databases conducted in Brazil, New Zealand, USA, and the UK.[15] Our accuracy results were similar to that described by other researchers for the linkage of primary data with the records of the SIM.[6] Predictive values, however, tend to be different in real situations because the prevalence of death was artificially increased in the study.

The question of representativeness of HAA records from the DATASUS was raised by the difficulty of finding HAA records in the accuracy tests and should be investigated further. Difficulties in identifying hospital admissions based on SIH has been described by other researchers.[4]

The probabilistic strategy used showed satisfactory accuracy and can be used in studies on the effectiveness of high-complexity and high-cost cardiology

procedures. The fifth blocking step resulted in excessive processing cost and provided a negligible contribution to the sensitivity of the method. The clerical review of pairs is the critical point of the process in terms of time, which indicates the need for greater systematization of this step and for studies that improve the method of calculating scores and increasing their powers of discrimination.

ACKNOWLEDGMENTS

We would like to thank Dr. Regina Elizabeth Müller, Instituto National de Cardiologia, for her contributions to the final review of the text and Maria Lucia Zurita Monteiro, Instituto National de Cardiologia, for her participation in the active searching process and in the final revision of the text.

REFERENCES

1. Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with Confidence. 2. ed. Bristol: British Medical Journal Books; 2000.
2. Camargo Jr KR, Coeli CM. Reclink: Aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage. *Cad Saude Publica*. 2000;16(2):439-47. DOI: 10.1590/S0102-311X2000000200014.
3. Camargo Jr KR, Coeli CM. Reclink 3: nova versão do programa que implementa a técnica de associação probabilística de registros (probabilistic record linkage). *Cad Saude Coletiva (Rio J)*. 2006;14(2):399-404.
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. Coeli CM, Camargo Jr KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. *Rev Bras Epidemiol*. 2002;5(2):185-96. DOI:10.1590/S1415-790X2002000200006
6. 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
7. Escosteguy CC, Portela MC, Medronho RA, Vasconcellos MTL. O Sistema de Informações Hospitalares e a assistência ao infarto agudo do miocárdio. *Rev Saude Publica*. 2002;36(4):491-9. DOI:10.1590/S0034-89102002000400016
8. Gordis L. Epidemiology. 4. ed. Philadelphia: Saunders Elsevier; 2009.
9. 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
10. Mathias TAF, Soboll MLMS. Confiabilidade de diagnósticos nos formulários de autorização de internação hospitalar. *Rev Saude Publica*. 1998;32(6):526-32. DOI:10.1590/S0034-89101998000600005
11. Melo ECP, Travassos C, Carvalho MS. Qualidade dos dados sobre óbitos por infarto agudo do miocárdio, Rio de Janeiro. *Rev Saude Publica*. 2004;38(3):385-91. DOI:10.1590/S0034-89102004000300008
12. Junger WL. Estimação de parâmetros em relacionamento probabilístico de bancos de dados: uma aplicação do algoritmo EM para o Reclink. *Cad Saude Colet (Rio J)*. 2006;14(2):225-32.
13. Silva JPL, Travassos C, Vasconcellos MM, Campos LM. Revisão sistemática sobre encadeamento ou linkage de bases de dados secundários para uso em pesquisa em saúde no Brasil. *Cad Saude Colet (Rio J)*. 2006;14(2):197-224.
14. Silva LK. Avaliação tecnológica e análise custo-efetividade em saúde: a incorporação de tecnologias e a produção de diretrizes clínicas para o SUS. *Cienc Saude Coletiva*. 2003;8(2):501-20. DOI:10.1590/S1413-81232003000200014
15. Silveira DP, Artmann E. Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática. *Rev Saude Publica*. 2009;43(5):875-82. DOI:10.1590/S0034-89102009005000060

This research was funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Departamento de Ciência e Tecnologia do Ministério da Saúde (Process No. 551402/2007-5).
This study was presented at the IX Brazilian Congress of Public Health, in Recife, Northeastern Brazil, in 2009.
The authors declare that there are no conflicts of interest.