Renata Gutierrez da Matta Coutinho[I]

Claudia Medina Coeli[II]

Eduardo Faerstein[III]

Dóra Chor[IV]

Sensitivity of probabilistic record linkage for reported birth identification: Pró-Saúde Study

ABSTRACT

The objective of the study was to evaluate the sensitivity of probabilistic record linkage for reported birth identification. Data from the Pró-Saúde Study cohort population were used comprising technical-administrative staff at a university in Rio de Janeiro, Brazil, in 1999. A total of 92 records of subjects were linked to the database of the Brazilian Information System on Live Births (SINASC) using RecLink II program. Both reduced and amplified strategies of clerical review were used. The sensitivity for birth identification with the reduced strategy was 60.9%, while with the amplified strategy was 72.8%. The limited number of fields available and the high proportion of homonymous names were major obstacles for the attainment of more accurate results.

DESCRIPTORS: Birth Registration. Diagnostic Techniques and Procedures. Sensitivity and Specificity. Information Systems. Vital Statistics.

INTRODUCTION

Database linkage has been used to monitor outcomes in cohort studies. It allows to joining data sets from different sources even when there is no univocal field identifier. Fields (e.g., name, date of birth) that are common to related databases are jointly used to estimate the probability that a given pair of records refers to the same individual.[1]

The accuracy of this probabilistic technique is affected by the number of fields available for comparison and quality of completion. When there are few fields available with low discriminatory power the likelihood of false-positive pairs is increased, i.e., although classified as true pairs they refer to different individuals. False-negative pairs are often originated from failures in either data collection or data entry.[1]

The accuracy of probabilistic record linkage is assessed by comparing the results obtained in the joining process with an independent source of information on outcomes of interest (gold standard). Since these sources are not easily available, few accuracy studies have been conducted.[2,5,6]

The objective of the present study was to assess the sensitivity of probabilistic record linkage in identifying births reported by female subjects in a longitudinal study.

METHODS

A cross-sectional study was conducted using probabilistic record linkage for identifying births reported by female subjects participating in the Pró-Saúde

[I] Programa de Pós-Graduação em Saúde Coletiva. Instituto de Medicina Social. Universidade do Estado do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

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

[III] Departamento de Epidemiologia. Instituto de Medicina Social. Universidade do Estado do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

[IV] Departamento de Epidemiologia e Métodos Quantitativos. Escola Nacional de Saúde Pública. Fundação Oswaldo Cruz. Rio de Janeiro, RJ, Brasil

Correspondence:
Renata Gutierrez da Matta Coutinho
Departamento de Epidemiologia - Instituto de Medicina Social
Universidade do Estado do Rio de Janeiro
R. São Francisco Xavier, 524
Pavilhão João Lyra Filho, 7º andar
Blocos D e E Maracanã
20550-900 Rio de Janeiro, RJ, Brasil
E-mail: retont@yahoo.com.br

Received: 8/30/2007
Revised: 6/3/2008
Approved: 6/10/2008

Study. The Brazilian Information System on Live Births (SINASC) database for the State of Rio de Janeiro was studied. Information on the date of birth of the first child of all subjects was the gold standard.

The Pró-Saúde Study is a longitudinal study including a sample of technical-administrative staff of a university in Rio de Janeiro.[3] For the present study there were selected female subjects included in phase 1 of the study data collection carried out in 1999 (n=2,238) who reported having their first liveborn child between 1996 and 1998 (n=92). SINASC database for the State of Rio de Janeiro, obtained from the State Health Department Office of Vital Statistics (1996 to 1998; N=798,478) provided identification information.

Linkage was performed using RecLink II software program.[1] A three-step blockage approach was applied based on a combination of phonetic codes of the fields "mother's last name" and "mother's first name". The fields used for pairing were "mother's name" and "mother's year of birth" (calculated based on the mother's age and date of birth).

All links with scores ≥0 were checked manually in the first step and only links with scores higher than six were manually reviewed in the next steps (short review). To improve the capture of true pairs, this strategy was expanded to manual review of all links with score ≥0 in all steps. During manual review the fields "mother's name," "mother's year of birth" and "district of residence" were checked.

Databases were evaluated for field completeness in automatic ("name" and "year of birth") and manual ("district of residence") processes. The sensitivity of probabilistic record linkage was calculated for both strategies of manual review for identification of records of births reported by the mothers. These estimates were repeated excluding births in the year 1996.

The study was approved by the Research Ethics Committee of Universidade do Estado do Rio de Janeiro Institute of Social Medicine.

RESULTS

The Pró-Saúde Study database showed 100% completion for subject's name and year of birth. As for SINASC database, an improvement was seen over the years studied in the field "name," which was completed in 73.6% of records in 1996, 90.5% in 1997, and 97.5% in 1998. For mother's year of birth, obtained from the field "age," completion rates were higher than 98% in all years studied.

The strategy of short manual review allowed to identifying 56 women (60.9% sensitivity; 95% CI: 50.7;70.2) out of 92 who reported having their first child between 1996 and 1998. The expanded strategy identified an additional 11 women, making a total of 67 (72.8% sensitivity; 95% CI: 63.0;80.9).

Due to inadequate completion of fields required to joining SINASC database in 1996, a sensitivity analysis was carried out excluding women who had their first child in that year. The sample total was then 63 women, of which 44 (69.8% sensitivity; 95% CI: 57.6;79.8) were identified through the short strategy and 55 (87.3% sensitivity; 95% CI: 76.9;93.4) through the expanded one.

DISCUSSION

The present study showed low sensitivity of the short strategy of manual review and moderate sensitivity of the expanded one. These results are less favorable than that reported in another Brazilian study of probabilistic linkage between a primary database (cohort of elderly patients admitted to hospitals due to fracture) and the Mortality Information System (SIM) for death identification, where 85% sensitivity was found.[2]

The high proportion of records with missing information on mother's name in SINASC database for the year 1996 can in part explain our results since the exclusion of that year from analysis increased sensitivity. However, a similar sensitivity to that reported by Coutinho & Coeli[2] was only achieved after applying the expanded manual review of links.

We noted that several links with high scores that were created for the same subject were mostly false pairs. The limited number of fields available for joining databases negatively affected the procedure's discriminatory power. Because childbearing women are part of close birth cohorts, it is common to find a high proportion of certain homonymous names "that are in fashion". Also, since a great number of Brazilians share common last names, a high rate of homonymous names with similar information on year of birth was found. An upper threshold score could not be determined, and false-positive pairs were even seen in links with the highest score (score=11). Thus, thorough manual review of links was required and strict criteria were established for final classification of pairs as true or false, which resulted in low sensitivity for birth identification in SINASC databases.

Linkage of SINASC and SIM databases to assess mortality in those under one year of age is an innovative application of this tool in Brazil.[4] This approach allows to join a varied set of fields with information on delivery and newborn that is available from both SINASC and SIM databases, which facilitates the linkage procedure. But, for joining SINASC database and other databases or for other purposes, as in the present study, this procedure is hindered by the limited number of fields and high proportion of homonymous names.

Although the expanded strategy provided adequate results for birth identification, this is a laborious approach. In the present study, one of the databases studied (Pró-Saúde Study) had a small number of records (n=92) but most applications involve large databases. For example, for joining all childbearing subjects in the Pró-Saúde Study (N=2.449) with SINASC database for a single year (≅ 270,000 records), an extremely large number of links would be required to be manually reviewed in the expanded strategy. While approximately 9,000 links would have to be reviewed in the first step, more than 200,000 links would have to be reviewed in the following steps.

In conclusion, the limited number of fields available and high proportion of homonymous names increased the probability of false-positive links, requiring manual review of a larger number of links and the establishment of strict criteria for final link classification. Our results suggest that linkage probabilistic procedure for joining SINASC databases for the purposes other than infant mortality assessment will have lower sensitivity than expected for joining databases from different sources in Brazil. The previous study of database completeness with exclusion of those years with high rates of missing information can contribute for more accurate results.

REFERENCES

1. 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
2. Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida. *Cad Saude Publica.* 2006 [cited 2007 Feb 15];22(10):2249-52. Available from: http://www.scielosp.org/pdf/csp/v22n10/24.pdf DOI: 10.1590/S0102-311X2006001000031
3. Faerstein E, Chor D, Lopes CS, Werneck GL. Estudo Pró-Saúde: características gerais e aspectos metodológicos. *Rev Bras Epidemiol.* 2005;8(4):454-66. DOI: 10.1590/S1415-790X2005000400014
4. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. *Cad Saude Publica.* 2004 [cited 2007 Feb 15];20(2):362-71. Available from: http://www.scielosp.org/pdf/csp/v20n2/03.pdf DOI: 10.1590/S0102-311X2004000200003
5. 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(1):54-7.
6. 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):1441-52. DOI: 10.1016/0895-4356(95)00530-7

Article based on RGM Coutinho's Master's dissertation submitted to Universidade do Estado do Rio de Janeiro in 2007.
RGM Coutinho was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes master's grant).
Supported by Conselho Nacional de Desenvolvimento Científico and Tecnológico (CNPq – Processes: 50. 0476/2004-7 CNPq Decree 03/2003 – Support to Technological and Scientific Research Projects and 47.1562 / 2004-1); CNPq Social Determinants Decree (409781/2006-1); and Fundação de Apoio à Pesquisa do Estado do Rio de Janeiro (Faperj – E-26/170.550/2004; E-26 / 100479/2007).