Rev Saúde Pública 2008;42(6):1097-100

Renata Gutierrez da Matta Coutinho[I]

Claudia Medina Coeli[II]

Eduardo Faerstein[III]

Dóra Chor[IV]

Sensibilidade do linkage probabilístico na identificação de nascimentos informados: Estudo Pró-Saúde

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

RESUMO

O objetivo do estudo foi avaliar a sensibilidade do método de linkage probabilístico de registros na identificação de nascimentos de coorte. Foram utilizados dados da população do Estudo Pró-Saúde, um estudo com funcionários técnico-administrativos do quadro efetivo de uma universidade no Rio de Janeiro, realizado em 1999. Os registros de 92 participantes foram relacionados com a base do Sistema de Informação sobre Nascidos Vivos utilizando o programa RecLink II. Empregaram-se estratégias de revisão manual reduzida e ampliada. A sensibilidade para a identificação dos nascimentos na estratégia reduzida foi de 60,9%, enquanto que na ampliada foi de 72,8%. Os poucos campos disponíveis e a elevada proporção de homônimas representaram os maiores obstáculos para a obtenção de resultados mais acurados.

DESCRITORES: Registro de Nascimento. Técnicas de Diagnóstico e Procedimentos. Sensibilidade e Especificidade. Sistemas de Informação. Estatísticas Vitais.

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.

[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

Correspondência | 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

Recebido: 30/8/2007
Revisado: 3/6/2008
Aprovado: 10/6/2008

INTRODUÇÃO

O *linkage* de bases de dados vem sendo utilizado para monitorar desfechos em estudos de coorte. Essa técnica permite integrar bases de dados de natureza diversa, mesmo na ausência de um campo identificador unívoco. Isso é possível por meio da utilização conjunta de campos comuns (e.g. nome, data de nascimento) às bases relacionadas, para estimar a probabilidade de que um par de registros se refira a um mesmo indivíduo.[1]

A acurácia do método probabilístico é influenciada pelo número de campos disponíveis para comparação e pela qualidade de preenchimento. Poucos campos disponíveis com baixo poder discriminatório aumentam a ocorrência de pares falso-positivos, i.e., apesar de classificados como pares verdadeiros, referem-se a pessoas diferentes. Os pares falso-negativos freqüentemente derivam de falhas na coleta da informação ou na digitação.[1]

Para a avaliação da acurácia do método de *linkage* probabilístico, é necessário comparar os resultados obtidos no processo de relacionamento com uma fonte de informação independente sobre a ocorrência dos desfechos de interesse (padrão-ouro). Como a disponibilidade dessas fontes é restrita, a realização de estudos de acurácia é reduzida.[2,5,6]

O objetivo do presente estudo foi avaliar a sensibilidade do método probabilístico de *linkage* na identificação de nascimentos informados por mulheres participantes de um estudo longitudinal.

MÉTODOS

Foi realizado um estudo seccional empregando-se o método de *linkage* probabilístico para a identificação de nascimentos informados pelas participantes do Estudo Pró-Saúde em uma base de dados do Sistema de Informação sobre Nascidos Vivos (SINASC) do Estado do Rio de Janeiro. A informação sobre a data de nascimento do primeiro filho das participantes foi utilizada como padrão-ouro.

O Estudo Pró-Saúde é um estudo longitudinal de funcionários técnico-administrativos do quadro efetivo de uma universidade no Rio de Janeiro.[3] Foram selecionadas para o presente estudo as mulheres participantes da fase 1 de coleta de dados do Estudo, realizada em 1999 (n=2.238), que afirmaram ter tido o primeiro filho nascido vivo entre 1996 e 1998 (n=92). A base de dados do SINASC do estado do Rio de Janeiro, obtida do Departamento de Dados Vitais da Secretaria Estadual de Saúde, (1996 a 1998; N=798.478) continha dados de identificação.

O *linkage* foi realizado por meio do programa *RecLink II.*[1] Empregou-se a estratégia de blocagem em três passos a partir da combinação de códigos fonéticos dos campos primeiro e último nome da mãe. Os campos usados no pareamento foram o nome e o ano de nascimento da mãe (calculado a partir da idade e data de nascimento da mãe).

Inicialmente foram revisados manualmente todos os *links* com escore ≥0 no primeiro passo e apenas os *links* com escores acima de seis nos demais passos (revisão reduzida). Para melhorar a captação de pares verdadeiros, essa estratégia foi ampliada para a revisão manual de todos os *links* com escore ≥0 em todos os passos. Durante o processo de revisão manual foram avaliados os campos nome, ano de nascimento da mãe e bairro de residência.

As bases de dados foram avaliadas em relação à completude dos campos empregados para os procedimentos automáticos (nome e ano de nascimento) e manual (bairro de residência). Calculou-se a sensibilidade do método de *linkage* probabilístico, nas duas estratégias de revisão manual, para a identificação dos registros de nascimento informados pelas mães. Esses cálculos foram repetidos excluindo-se os nascimentos ocorridos no ano de 1996.

O estudo foi aprovado pelo Comitê de Ética em Pesquisa do Instituto de Medicina Social da Universidade do Estado do Rio de Janeiro.

RESULTADOS

A base de dados do Estudo Pró-Saúde apresentou preenchimento de 100% para o nome e ano de nascimento da participante. Com relação ao SINASC, houve melhora ao longo dos anos estudados para o campo nome, que apresentou preenchimento em 73,6%, dos registros em 1996, de 90,5% em 1997 e de 97,5%, em 1998. Para o ano de nascimento da mãe, derivado do campo idade, os valores foram superiores a 98% em todos os anos estudados.

Empregando-se a estratégia de revisão manual reduzida, foi possível a identificação de 56 mulheres (sensibilidade = 60,9%; IC 95%: 50,7;70,2%) das 92 que afirmaram ter tido o primeiro filho entre 1996-1998. Ao utilizar a estratégia ampliada, foram identificadas mais 11 mulheres, totalizando 67 (sensibilidade = 72,8%; IC 95%: 63,0;80,9%).

Devido ao preenchimento insuficiente dos campos necessários ao relacionamento no banco de dados do SINASC em 1996, foi feita análise da sensibilidade excluindo as mulheres que tiveram o primeiro filho nesse ano. O total da amostra foi alterado para 63, das quais 44 mulheres (sensibilidade = 69,8%; IC 95%: 57,6;79,8%) foram identificadas pela estratégia inicial e 55 (sensibilidade= 87,3%; IC 95%: 76,9;93,4%) por meio da estratégia ampliada.

DISCUSSÃO

Os resultados do presente estudo mostraram uma sensibilidade baixa quando da aplicação da primeira estratégia de revisão manual, e moderada quando se utilizou a estratégia ampliada. Esse resultado foi menos favorável do que o observado em outro estudo desenvolvido no Brasil, em que foi realizado o *linkage* probabilístico de uma base de dados primários (coorte de idosos hospitalizados por fratura) e o Sistema de Informação Sobre Mortalidade (SIM) para a identificação de óbitos, sendo observada uma sensibilidade de 85% para o método de *linkage*.[2]

O elevado percentual de registros sem informação no SINASC sobre o nome da mãe no ano de 1996 pode explicar parcialmente nosso resultado, já que a exclusão desse ano da análise aumentou os valores de sensibilidade. Entretanto, sensibilidade semelhante à observada por Coutinho & Coeli[2] só foi alcançada em nosso estudo após aplicação de estratégia ampliada de revisão manual de *links*.

Observamos a formação de vários *links* com escore elevado para uma mesma participante, que em sua maioria representaram pares falsos. O reduzido número de campos disponíveis para o relacionamento das bases de dados diminuiu o poder discriminatório do método. Além disso, mulheres em idade reprodutiva pertencem a coortes de nascimento próximas, sendo comum observar uma maior proporção de homônimas em função de determinados nomes "da moda". Como no Brasil também é observada uma grande concentração de determinados sobrenomes, foi observada uma grande proporção de homônimas com informação similar sobre o ano de nascimento. Não foi possível identificar um valor de escore limiar superior, sendo observados pares falso-positivos mesmo entre os *links* que alcançaram o valor de escore máximo (escore=11). Assim, foi necessário realizar uma cuidadosa revisão manual dos *links* e aplicar critérios rígidos para a classificação final dos pares em verdadeiros ou falsos, o que levou a uma perda de sensibilidade para a identificação de nascimentos nas bases do SINASC.

O *linkage* das bases de dados do SINASC e do SIM, visando avaliar a mortalidade em menores de um ano, vem representando uma das aplicações pioneiras de estratégias de *linkage* no Brasil.[4] Nesse tipo de exercício podem ser utilizados um conjunto variado de campos com informações sobre o parto e o recém-nascido presentes tanto no SINASC, como no SIM, o que facilita o processo de *linkage*. Para aplicações envolvendo o relacionamento do SINASC com outras bases e outros objetivos, como foi o caso do presente estudo, o processo é dificultado pelo reduzido número de campos e elevada proporção de homônimos.

Embora a estratégia ampliada empregada tenha alcançado resultados satisfatórios para a identificação de nascimentos, trata-se de procedimento trabalhoso. No presente estudo, uma das bases de dados (Estudo Pró-Saúde) apresentava um número reduzido de registros (n=92). Entretanto, a maioria das aplicações envolve bases de maior tamanho; por exemplo, para relacionar a base de todas as participantes em idade reprodutiva do Estudo Pró-saúde (N=2.449) com a base de um único ano do SINASC (≅ 270.000 registros) a aplicação da estratégia ampliada envolveria número excessivo de *links* demandando revisão manual. Enquanto para o primeiro passo seria necessário rever aproximadamente 9.000 *links*, para os demais seria necessária a revisão de mais de 200.000 *links* em cada passo.

Concluindo, o reduzido número de campos disponíveis e a elevada proporção de homônimas aumentou a probabilidade de ocorrência de *links* falso-positivos, levando à necessidade da revisão manual de um número maior de *links* e o emprego de regras restritas para a classificação final do *link* como par verdadeiro. Nossos resultados sugerem, portanto, que aplicações da metodologia de *linkage* probabilístico envolvendo o relacionamento das bases do SINASC para objetivos que não visem à avaliação da mortalidade infantil, devem apresentar sensibilidade menor que a esperada para o relacionamento entre bases de outra natureza no Brasil. O estudo prévio da completude das bases com a exclusão dos anos com elevado percentual de não-preenchimento de campos pode contribuir para a obtenção de resultados mais acurados.

REFERÊNCIAS

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[citado 2007 fev 15];22(10):2249-52. Disponível em: 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[citado 2007 fev 15]; 20(2):362-71. Disponível em: 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

Artigo baseado na dissertação de mestrado de RGM Coutinho, apresentada à Universidade do Estado do Rio de Janeiro, em 2007.
RGM Coutinho foi apoiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes – bolsa de mestrado).
Financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – Processos: 50. 0476/2004-7 Edital CNPq 03/2003 – Apoio Técnico a Projetos de Pesquisa Científica e Tecnológica e 47.1562 / 2004-1); pelo Edital Determinantes Sociais CNPq (409781/2006-1); e pela Fundação de Apoio à Pesquisa do Estado do Rio de Janeiro (Faperj – E-26/170.550/2004; E-26 / 100479/2007).