Notas e Informações
Notes and Informations

Vinculação Determinística de Bancos de Dados sobre Mortalidade por Aids

Deterministic record linkage in Aids mortality databases

Bruna Bronhara[1]
Wolney Lisboa Conde[1]
Daniele Carli Liciardi[2]
Ivan França-Junior[2]

[1] Departamento de Nutrição da Faculdade de Saúde Pública da Universidade de São Paulo

[2] Departamento de Saúde Materno Infantil da Faculdade de Saúde Pública da Universidade de São Paulo.

Agências de fomento: Este estudo deriva de dois projetos: "Estigma e Discriminação relacionados ao HIV/AIDS: impactos da epidemia em crianças e jovens na cidade de São Paulo", (Processos FAPESP nº 2003/10883-5, 2005/59566-7 e 06/50553-2) e "Impactos das mortes por homicídio e aids na saúde e nos direitos de crianças e jovens na cidade de São Paulo" (Processo CNPq n° 476210/2004-6).
Correspondência: Ivan França Junior Departamento de Saúde Materno Infantil da Faculdade de Saúde Pública - USP, Avenida Dr. Arnaldo, 715 CEP 01246-904 - São Paulo, SP. E-mail: ifjunior@usp.br

Resumo

A vinculação determinística de bancos de dados sobre mortalidade por aids tem apresentado problemas causados por falhas nos arquivos. Assim, os objetivos deste estudo foram: avaliar o desempenho da vinculação determinística em bancos de óbito por aids do Programa de Aprimoramento das Informações de Mortalidade no Município de São Paulo (PRO-AIM) e da Fundação SEADE entre os anos de 2000 e 2004 e estimar a cobertura de cada banco. Utilizou-se a rotina *merge* de um *software* para vincular os bancos. A primeira etapa pareou os registros automaticamente e, na segunda etapa, cada banco foi conferido para localizar novos pares. Estimaram-se os óbitos pela soma entre casos pareados e não pareados para calcular a cobertura dos bancos. A primeira etapa da vinculação identificou 91,6% dos pares. A segunda etapa adicionou 457 pares. O total de óbitos foi estimado em 5.855, com cobertura de 97,1% do PRO-AIM e 96% do SEADE. O uso da vinculação determinística cobriu grande parte dos casos. O banco do PRO-AIM proporcionou a maior cobertura, com maior quantidade de informações completas e melhor localização geográfica dos casos.

Palavras-chave: Sistemas de Informações sobre Mortalidade. Aids. Registros de Mortalidade.

Abstract

Deterministic record linkage in AIDS mortality databases has presented problems caused by errors in files. Thus, the aims of this study were to assess the performance of deterministic record linkage in the Aids mortality databases of the Mortality Information Improvement Program of the city of Sao Paulo (PRO-AIM) and of the SEADE Foundation between the years 2000 and 2004 and to estimate the coverage of these databases. A software merge process was used to link the records. The first stage linked automatically and, in the second stage, each database was checked for possible matches. Deaths were estimated by summing the matches and non-matches for each database coverage calculation. The first stage of record linkage identified 91.6% matches. The second stage added 457 matches. Deaths were estimated to be 5,855 cases, PRO-AIM covering 97.1% and SEADE 96%. The use of deterministic record linkage covered a great part of the cases. PRO-AIM databases provided the best case coverage, with the largest amount of complete information and better geographic location.

Key Words: Mortality Information System. AIDS. Mortality Registers.

Introdução

No município de São Paulo, há a possibilidade de se examinar diferenças entre dois sistemas de informações sobre mortalidade por aids, operados por dois órgãos públicos (Programa de Aprimoramento das Informações de Mortalidade no Município de São Paulo / PRO-AIM[1] e Fundação Sistema Estadual de Análise de Dados / Seade[2]). Estas instituições coletam dados das mesmas declarações de óbitos (DO). O PRO-AIM, adicionalmente, realiza busca ativa em DO com referências à aids e outras doenças de notificação compulsória, notificando os órgãos de vigilância epidemiológica e solicitando esclarecimentos aos médicos declarantes no caso de causa básica de morte suspeita[3]. A vinculação entre estes dois bancos de dados permite aferir a qualidade dos sistemas de informação em mortalidade e agregar outras informações sobre o indivíduo, possibilitando estudos com baixo custo operacional[4,5].

Os métodos de vinculação de bancos podem ser determinísticos ou probabilísticos. O procedimento probabilístico envolve a ponderação das informações utilizadas para a vinculação dos registros. A maioria dos *softwares* utilizados em estudos epidemiológicos não dispõe de rotinas automatizadas para tal procedimento. A vinculação determinística é um método que associa os registros segundo identificadores únicos[4], os quais devem ser idênticos para que os dados sejam vinculados automaticamente. Entretanto, boa parte do insucesso no pareamento destes bancos se deve à ausência de preenchimento dos campos, falhas na codificação, digitação dos dados ou estrutura dos bancos.

Assim, esse trabalho tem como objetivos avaliar o desempenho da vinculação determinística, convencional e modificada, em bancos de óbitos por aids provenientes do PRO-AIM[1] e da Fundação SEADE[2]; e estimar a cobertura municipal da informação sobre mortalidade por aids.

Métodos

Utilizaram-se os bancos de óbitos codificados como causa básica aids durante o período de 2000 a 2004, fornecidos pelo PRO-AIM[1] e pela Fundação SEADE[2]. Selecionaram-se apenas os registros em que o município de residência e de ocorrência do óbito correspondiam ao município de São Paulo e os que apresentavam idade maior ou igual a 18 anos no momento do óbito (PRO-AIM: *n*=5687 e SEADE: *n*=5625). Para vincular esses dois bancos, foi utilizada a rotina *merge* do programa *Stata 9,2*[6].

A vinculação foi realizada em duas etapas. Na primeira etapa, foram selecionadas as seguintes informações: nome, sexo, data de nascimento e de óbito e adotados três passos para pareamento automático dos registros. No primeiro passo, a informação contida no campo nome foi transformada em variável construída com letras maiúsculas e sem espaços entre letras e palavras, identificada como *nome1*. Por exemplo, o nome "Maria da Silva" transformou-se em "MARIADASILVA". Para realização do pareamento no passo 1, as variáveis selecionadas foram *nome1, sexo, data de nascimento completa* e *data de óbito completa*. Após essa etapa, os registros não pareados de cada banco foram submetidos a novo tratamento. Nesta segunda etapa, a variável *data de nascimento completa* foi dividida em três novas variáveis: *dia de nascimento, mês de nascimento e ano de nascimento*. As variáveis selecionadas para esta etapa foram *nome1, sexo, mês de nascimento, ano de nascimento* e *data de óbito completa;* os registros não pareados foram submetidos ao passo 3. Na terceira etapa, o primeiro e o segundo nomes do registro foram condensados na variável *nome2*. Foram utilizadas ainda as datas de nascimento e óbito para o pareamento.

Outras combinações de variáveis também foram testadas, produzindo, porém, poucos casos adicionais e, mais frequentemente, pareamentos errados. Assim, o passo 3 foi considerado o limite máximo para o pareamento convencional.

Na segunda etapa, o pareamento modificado iniciou-se com a busca manual de casos com nomes semelhantes para eventuais correções ortográficas das variáveis *nome*, *sexo, data de nascimento* e *data de óbito* (passo 4). A seguir, os registros dos dois bancos foram submetidos à rotina *merge,* conforme descrito no passo 1.

O total de óbitos por Aids ocorridos entre 2000-2004 foi estimado pela soma dos casos pareados e não pareados dos bancos PRO-AIM e SEADE. A partir do total estimado, calculou-se a contribuição de cada um dos sistemas de informação.

Resultados

A primeira etapa da vinculação determinística identificou 91,6% (*n*=5000) do total de pares formados, correspondendo a 87,9% do total de registros do PRO-AIM e 88,8% do SEADE. Os passos 2 e 3 pouco acrescentaram ao total de casos pareados nesta etapa (3,1%).

A segunda etapa adicionou 457 pares, chegando a 96% do PRO-AIM e 97% do SEADE (Tabela 1). Estes 457 pareamentos representaram 8,3% do total de pareados.

O total de falecimentos entre 2000-2004 codificados como causa básica aids, no município de São Paulo e de indivíduos com idade maior ou igual a 18 anos, foi estimado em 5.855 casos. A contribuição dos sistemas de informação para a cobertura total foi de 97,1% do PRO-AIM e 96% do SEADE em relação ao total estimado.

No conjunto de registros não pareados do PROAIM, todas as informações dos campos *sexo* e *data de óbito* estavam preenchidas. Os campos *idade* e *endereço* apresentaram 3% (*n*=7) e 37% (*n*=85) de informações ignoradas, respectivamente e, em *endereço,* também foram encontrados 3% (*n*=7) de informações em branco.

Em relação ao SEADE, todos os registros dos campos ao *sexo, idade, data de nascimento* e *data de óbito* estavam preenchidos. O campo *endereço* apresentou 87 casos (51%) inutilizáveis, por informações ignoradas (*n*=3) ou em branco (*n*=84).

Tabela 1 – Número de registros pareados e não pareados em cada passo da vinculação determinística.
Table 1 – *Number of matching and non-matching records in each stage of deterministic linkage.*

Registro	Primeira Etapa		Segunda Etapa		Total
	Passo 1	Passo 2	Passo 3	Passo 4	
Pareados (*n*)	4846	36	118	457	5457
Não Pareados (*n*)					
PRO-AIM	841	805	687	230	230
Fundação SEADE	779	743	625	168	168
Total					5855

Fonte/*Source*: Fundação SEADE e PROAIM.
Passo 1 = nome*, sexo, data de nascimento e data de óbito. *Step 1 = name*, gender, date of birth, and date of death.*
Passo 2 = nome*, sexo, mês de nascimento, ano de nascimento e data de óbito. *Step 2 = name*, gender, month of birth, year of birth, and date of death.*
Passo 3 = primeiro + segundo nomes, data de nascimento e data de óbito. *Step 3 = first + second names, date of birth, and date of death.*
Passo 4 = correção dos bancos e posterior passo 1. *Step 4 = correction of databases and then step 1.*
*nome: variável *nome* modificada com letras maiúsculas e sem espaços entre letras e palavras. **name: variable* name *modified with lowercase and without spaces between letters and words.*

Discussão

No presente estudo, o uso de procedimentos determinísticos para vinculação dos bancos sobre mortalidade por aids preencheu amplo espectro dos casos disponíveis. A elevada freqüência de atributos, como nome, data de nascimento e data de óbito, passíveis de utilização como identificadores, torna o procedimento determinístico preferível inicialmente ao probabilístico, particularmente naqueles casos que possam ser inspecionados manualmente.

A vinculação probabilística consiste no pareamento dos registros com base em escores construídos a partir do emprego de determinada estratégia de blocagem[7]. Este processo, segundo Camargo Jr & Coeli[7] consiste na "indexação dos arquivos a serem relacionados segundo uma chave formada por um campo ou pela combinação de mais de um campo". A maioria dos *softwares* utilizados em estudos epidemiológicos não apresenta rotinas automatizadas para tais procedimentos. Assim, o uso de procedimentos determinísticos permitiria a maior difusão da utilização de bancos de dados dos sistemas públicos de informação sobre mortalidade.

A busca manual dos registros semelhantes evitou a presença de registros duplicados ao estimar os casos de óbitos por Aids, impossibilitando a superestimação do total de falecimentos. Esse procedimento metodológico foi também adotado para vincular dois bancos oficiais de informação de mortalidade e nascidos vivos em estudo sobre mortalidade neonatal[5].

A vinculação determinística de bancos foi também utilizada por Silva et al.[8], para verificar a concordância entre as informações constantes no Sistema de Informação de Nascidos Vivos referentes a partos hospitalares e aquelas obtidas em inquérito epidemiológico. Os autores vincularam 72,4% dos registros automaticamente. Pelo processo manual, identificaram 1,2% de registros pareados incorretamente e mais de 4,6% registros semelhantes, o que proporcionou 75,8% de sucesso na vinculação determinística.

A análise dos registros não pareados em cada banco indicou que o PRO-AIM, apesar de apresentar variáveis com informações perdidas, ainda possui a maior quantidade de informações sobre os casos e possibilita a melhor localização geográfica dos mesmos, possivelmente devido ao procedimento de busca ativa realizada a partir das declarações de óbito a qual resulta em melhor informação.

Conclusão

O uso da vinculação determinística cobriu grande parte dos casos analisados

e pode ser melhorado pela correção de erros de ortografia/digitação. O banco proveniente do PRO-AIM foi responsável pela maior cobertura dos casos estimados de óbitos por Aids ocorridos no período. A análise dos registros não pareados de cada banco mostrou também que o PRO-AIM possui maior quantidade de informações sobre os casos e possibilita a melhor localização geográfica dos mesmos. Estudos sobre a eficiência do uso da vinculação determinística em outros sistemas de informação são necessários para se conhecer mais amplamente o potencial deste recurso de referência no manejo de informações epidemiológicas.

Referências

1. Secretaria Municipal de Saúde [CD ROM]. Programa de Aprimoramento de Informações sobre mortalidade; 2006.

2. Fundação SEADE. Demografia do Município de São Paulo**.** http://www.seade.gov.br/produtos/msp/menu_tema_4.php?opt=s&tema=DEM&subtema=5 [Acessado em 14 de outubro de 2004].

3. Drumond Jr M, Lira MMTA, Freitas M, Nitrini TMV. A AIDS e os sistemas de informação de mortalidade em nível local: A experiência do PROAIM no Município de São Paulo. *AIDS - Boletim Epidemiológico* 1997; 9: 3-9.

4. Machado CJ. Procedimentos para relacionamento de registros: revisão bibliográfica com enfoque na saúde infantil. *Cad Saúde Pública* 2001; 20(2): 362-71.

5. Almeida MF, Mello-Jorge MHP. O uso da técnica linkage de sistemas de informação em estudos de coorte sobre mortalidade neonatal. *Rev Saúde Pública*, 30(2); 141-7, 1996.

6. StataCorp. Stata Statistical Software: release 8.0. College Station: Stata Corporation; 2003.

7. 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(2): 439-47.

8. Silva A et al. Avaliação da qualidade dos dados do Sistema de Informações sobre Nascidos Vivos em 1997-1998. *Rev Saúde Pública* 2001; 35(6); 508-14.

Recebido em: 19/02/08
Versão final reapresentada em: 23/06/08
Aprovado em: 21/07/08