Estimativas de parâmetros no *linkage* entre os bancos de mortalidade e de hospitalização, segundo a qualidade do registro da causa básica do óbito

Estimated parameters in linkage between mortality and hospitalization databases according to quality of records on underlying cause of death

Cláudia Medina Coeli [1]
Flávia dos Santos Barbosa [2]
Alexandre dos Santos Brito [3]
Rejane Sobrino Pinheiro [1]
Kenneth Rochel de Camargo Jr. [2]
Roberto de Andrade Medronho [1]
Katia Vergetti Bloch [1]

[1] Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.
[2] Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.
[3] Departamento de Epidemiologia e Bioestatística, Universidade Federal Fluminense, Niterói, Brasil.

Correspondência
C. M. Coeli
Instituto de Estudos em Saúde Coletiva, Universidade Federal do Rio de Janeiro.
Pça. Jorge Moreira Machado 1000, Rio de Janeiro, RJ 21941-598, Brasil.
coeli@iesc.ufrj.br

Abstract

The purpose of the study was to compare the linkage parameter estimates between hospitalization and mortality databases, calculated separately for the subsets of deaths from ill-defined causes and deaths from known causes. The databases for deaths from known causes and ill-defined causes were linked to a hospital admissions database. Parameters were estimated using two strategies: (1) first name, last name, and day, month, and year of birth, (2) full name and date of birth. In the first strategy, the estimates for the first and last name were at least 97% in both sets. However, the items day, month, and year of birth produced low values in both sets. In the second strategy there was an important difference between the two groups, with much lower values for full name and especially for date of birth in the group of deaths from ill-defined causes. Our results emphasize the need for pilot studies to evaluate possible internal heterogeneity of databases during the planning stage of linkage projects.

Cause of Death; Mortality; Database

Introdução

O interesse no uso de dados secundários na pesquisa e avaliação em saúde vem aumentando com a crescente disponibilidade de grandes bases de dados e de programas computacionais que viabilizam o uso integrado deles [1,2,3].

A integração de bases de dados de naturezas diversas traz como vantagem a possibilidade de ampliação do escopo de hipóteses que podem ser testadas. Entretanto, para que resultados válidos possam ser obtidos é fundamental que o processo de integração se dê com a ocorrência mínima de erros [3,4].

No Brasil, não existe um identificador unívoco nas bases de dados disponíveis, sendo necessário utilizar técnicas de *linkage* probabilísticas. O modelo desenvolvido por Fellegi & Sunter [5], um dos mais empregados, baseia-se na utilização de campos identificadores comuns presentes nas bases. Esses identificadores são usados conjuntamente para o cálculo de um escore que traduz o grau de concordância entre os registros de cada *link* formado [2,3].

Para cada campo *i* define-se a probabilidade m_i do campo concordar entre os dois registros, dado que se trata de um par verdadeiro, e a probabilidade u_i do campo concordar, dado que se trata de um par falso. Tais probabilidades são os parâmetros de *linkage* do modelo usadas para a construção de dois pesos (concordância e discor-

dância). Compara-se o campo do primeiro registro com o do segundo e, se os campos concordarem, aplica-se o peso de ponderação de concordância e, no caso contrário, o de discordância. O peso de concordância é calculado como o logaritmo de base 2 da razão de verossimilhanças entre m_i e u_i, e o de discordância como o logaritmo de base 2 da razão de verossimilhanças entre 1-m_i e 1-u_i. O peso total de um determinado *link* é obtido pela soma dos pesos parciais atribuídos após a comparação de cada campo avaliado [2,3]. Quanto maior for m_i e menor for u_i, maior será a contribuição do campo para o peso total do *link* e mais discriminatório o campo será.

A despeito do uso de algoritmos robustos para a comparação de campos, erros de informação e tipográficos impactam negativamente os parâmetros, especialmente o m_i. Dessa forma, esses parâmetros deveriam ser estimados a cada projeto de *linkage*, já que a qualidade das informações pode variar segundo a natureza das bases a serem relacionadas, o período e a localização geográfica. Em algumas situações, são esperadas variações mesmo tendo em conta que os aspectos acima listados sejam fixos. Por exemplo, espera-se que registros de óbito com causas mal definidas apresentem pior qualidade no preenchimento dos campos de identificação, quando comparados aos com causas definidas. O objetivo do presente estudo foi avaliar essa questão, comparando as estimativas de parâmetros de relacionamento entre as bases das Autorizações de Internação Hospitalar (AIH) e das Declarações de Óbito (DO) calculadas separadamente para os subconjuntos de óbitos com causas mal definidas e óbitos com causas definidas.

Métodos

Foram utilizadas as bases das DO relativas ao Estado do Rio de Janeiro no ano de 2001 (N = 116.000) e o banco com registros de todas as AIH do tipo 1 no mesmo período (N = 810.397). Com os dados do banco de DO de 2001, foram selecionados dois conjuntos de registros: um com cerca de 12.000 óbitos com causas mal definidas e outro composto por uma amostra aleatória simples, de mesmo tamanho, de óbitos com causas definidas. O banco das AIH analisado foi composto apenas com as internações que terminaram em óbito (N = 35.418).

O *linkage* probabilístico foi realizado pelo emprego do programa RecLink III [6]. Foi feita a padronização das bases e a quebra dos campos em seus componentes. As estimativas foram feitas considerando duas estratégias de comparação: (1) primeiro nome, último nome, dia, mês e ano de nascimento; (2) nome completo e data do nascimento. Os campos nome completo, primeiro e último nome foram comparados mediante uso de algoritmos baseados na distância de Levenshtein. O campo data de nascimento foi avaliado pela utilização dum algoritmo que compara caractere a caractere segundo sua posição. Já os campos dia, mês e ano de nascimento foram comparados empregando-se um algoritmo para a diferença de valor (+/-2). Os parâmetros m_i de relacionamento foram estimados pelo uso da chave de blocagem formada pela combinação entre o código *soundex* do primeiro e último nome e o sexo, sendo utilizada rotina baseada em algoritmo EM [7]. Já para o parâmetro u_i, o programa RecLink III cria uma matriz teórica com todos os *links* possíveis de serem obtidos sem blocagem, sendo, então, extraída uma amostra aleatória deles e estimada a frequência de concordância ao acaso.

As bases foram obtidas na Secretaria de Estado de Saúde e Defesa Civil do Rio de Janeiro, após aprovação do projeto pelo Comitê de Ética em Pesquisa do Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro (nº. 37/2007).

Resultados

Na primeira estratégia, as estimativas do parâmetro m_i para os campos primeiro e último nome foram iguais ou superiores a 97% em ambos os conjuntos analisados (Tabela 1). As estimativas foram semelhantes para ambos os conjuntos de dados para dia e mês, sendo observada diferença importante apenas para o campo ano, que apresentou pior estimativa no conjunto de óbitos com causas mal definidas. Na segunda estratégia houve diferença importante entre os dois grupos, com valores de m_i bem menores para nome completo e, especialmente, data de nascimento no grupo dos óbitos com causas mal definidas. As estimativas de u_i foram iguais para os dois subconjuntos estudados, exceto para o ano de nascimento.

Os menores valores de m_i estimados no subconjunto de óbitos com causas mal definidas para ano de nascimento, na primeira estratégia, e nome completo e data de nascimento, na segunda, implicaram menor poder discriminatório dos campos de comparação nesse estrato, o que pode ser evidenciado por valores absolutos menores de pesos de concordância e discordância no subconjunto de óbitos (Tabela 1).

Tabela 1

Probabilidades m_i e u_i, escores individuais para concordância e discordância no *linkage* dos registros de Autorização de Internação Hospitalar (AIH) e registros de óbitos com causa definida e mal definida.

Estratégia/Campo	Probabilidade (m_i)		Probabilidade (u_i)		Escore para concordância [Log_2 (m/u)]		Escore para discordância [Log_2 {(1-m)/(1-u)}]	
	Causas definidas	Causas mal definidas	Causas definidas	Causas mal definidas	Causas definidas	Causas mal definidas	Causas definidas	Causas mal definidas
Primeira estratégia								
Primeiro nome	0,97	0,98	0,02	0,02	5,88	5,83	-5,16	-5,99
Último nome	0,99	0,99	0,03	0,03	5,12	5,16	-6,17	-7,22
Dia de nascimento	0,57	0,42	0,15	0,15	1,94	1,44	-1,00	-0,54
Mês de nascimento	0,76	0,73	0,36	0,36	1,07	1,01	-1,41	-1,28
Ano de nascimento	0,47	0,24	0,07	0,08	2,68	1,06	-0.81	-0,27
Segunda estratégia								
Nome completo	0,91	0,61	0,00002	0,0002	15,67	14,7	-3,55	-1,37
Data de nascimento	0,92	0,00001	0,00002	0,0004	5,73	-1,88	-3,65	0,0005

Discussão

Nossos resultados confirmaram a hipótese inicial de que uma pior qualidade do registro dos dados de identificação no subconjunto de óbitos com causas mal definidas poderia levar a valores mais baixos das estimativas de m_i e, consequentemente, a um menor poder discriminatório dos campos disponíveis para os processos de *linkage*. O fato de as estimativas de m_i para nome completo serem mais baixas para os óbitos com causas mal definidas, enquanto as estimativas de primeiro e último nomes serem semelhantes nos dois conjuntos, poderia ser explicado por uma proporção maior de registros apresentando nomes abreviados no conjunto dos óbitos com causas mal definidas. Os resultados mostram um registro de pior qualidade da data de nascimento no grupo dos óbitos com causas mal definidas, entretanto as estimativas de m_i para dia, mês e ano apresentaram resultados ruins também no grupo dos óbitos com causas definidas. Esse achado sugere que não apenas erros de informação, mas também erros tipográficos possam ter provocado inconsistências nos registros da data de nascimento em ambas as bases. Winkler [8] observou diferenças importantes das estimativas de m_i entre regiões urbanas e suburbanas adjacentes nas bases de censo dos Estados Unidos, tendo atribuído esse achado a variações na ocorrência de erros tipográficos.

A etapa final do processo de *linkage* implica a classificação dos *links* para a identificação de pares verdadeiros. Caso fosse possível conhecer o *status* verdadeiro de cada *link*, poderiam ser construídas duas curvas relativas às distribuições de pesos totais dos pares falsos e dos pares verdadeiros. Como essas curvas sempre apresentam certo grau de superposição, o desafio é estimar, para um nível de erro conhecido, dois limiares de pesos que permitam a classificação dos *links* em pares verdadeiros (peso acima do limiar superior), falsos (peso abaixo do limiar inferior) e duvidosos (peso entre os dois limiares). Esses últimos podem ser encaminhados para a revisão manual com vistas à classificação final [2,3]. Várias metodologias são sugeridas para a classificação dos *links* [8], contudo todas elas são influenciadas pelo poder discriminatório do processo de *linkage*. Processos que empregam muitos campos identificadores com boa qualidade de preenchimento tendem a gerar maior separação das curvas, facilitando a classificação final dos pares. Quanto menos discriminatório for o processo, maior necessidade de revisão manual e menor acurácia são esperadas.

Em um estudo com *linkage* de óbitos com causas mal definidas e óbitos com causas definidas com registros de AIH relativos às saídas por óbito, Teixeira et al. [9] encontraram quatro vezes mais pares relacionados entre os óbitos com causas definidas, comparativamente aos óbitos com causas mal definidas. Nesse estudo foram empregados para comparação o nome completo e a data de nascimento e utilizados para os dois grupos os mesmo valores de parâmetros provenientes da literatura. Um menor número de pares formados entre os registros de óbitos com causas mal definidas é coerente com a esperada menor cobertura de atenção médico-hospitalar

e, por essa razão, menor probabilidade de causa de óbito identificada nesse grupo. Todavia, diferenças em erros de *linkage* causadas por heterogeneidades na qualidade dos identificadores nos dois subconjuntos de óbitos também podem justificar as diferenças encontradas. Nossos resultados indicam que a adoção de uma estratégia envolvendo a comparação de componentes do nome e da data de nascimento, assim como a estimativa de parâmetros em separado para os dois grupos, poderia melhorar a sensibilidade do processo de *linkage*.

Nossos resultados reforçam a necessidade de serem realizados estudos pilotos para avaliar possíveis heterogeneidades da qualidade de dados internas das bases durante o planejamento de projetos de *linkage,* permitindo a estimativa de parâmetros e a identificação da estratégia mais efetiva aos objetivos analíticos almejados.

Resumo

O objetivo do estudo foi comparar as estimativas de parâmetros de linkage *entre as bases de hospitalizações e de óbitos calculadas separadamente para uma amostra de óbitos com causas mal definidas e para os com causas definidas. As estimativas de parâmetros foram feitas considerando duas estratégias: (1) primeiro nome, último nome, dia, mês e ano de nascimento; (2) nome completo e data do nascimento. Na primeira estratégia, as estimativas do parâmetro para os campos primeiro e último nome foram iguais ou superiores a 97% em ambos os conjuntos analisados. Já os campos dia, mês e ano apresentaram valores baixos em ambos os conjuntos. Na segunda estratégia, houve diferença importante entre os dois grupos, com valores bem menores para nome completo e, especialmente, data de nascimento no grupo dos óbitos com causas mal definidas. Nossos resultados reforçam a necessidade de serem realizados estudos pilotos para avaliar possíveis heterogeneidades internas nas bases durante a fase de planejamento de projetos de* linkage.

Causas de Morte; Mortalidade; Base de Dados

Colaboradores

C. M. Coeli participou da escolha do tema, planejamento do estudo, análise de dados e redação do texto. F. S. Barbosa colaborou no planejamento do estudo, processamento e análise de dados e redação do texto. A. S. Brito contribuiu na análise e processamento de dados, revisão e edição final do texto. R. S. Pinheiro participou da coleta de dados, interpretação de resultados, revisão e edição final do texto. K. R. Camargo Jr. e R. A. Medronho colaboraram nas etapas de interpretação de resultados, revisão e edição final do texto. K. V. Bloch participou da escolha do tema, coleta de dados, interpretação de resultados, revisão e edição final do texto.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ) pelo apoio financeiro.

Referências

1. Pinheiro RS, Camargo Jr. KR, Coeli CM. Relacionamento de bases de dados em saúde. Cad Saúde Colet (Rio J.) 2006;14:195-6.
2. Camargo Jr. KR, Coeli CM. Reclink: aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage. Cad Saúde Pública 2000; 16:439-47.
3. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques. New York: Springer; 2007.
4. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. Stat Med 1997; 16:2633-43.
5. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc 1969; 64:1183-210.
6. Camargo Jr. KR, Coeli CM. RecLink 3: nova versão do programa que implementa a técnica de associação probabilística de registros (probabilistic record linkage). Cad Saúde Colet (Rio J.) 2006; 14:399-404.
7. Junger WL. Estimação de parâmetros em relacionamento probabilístico de banco de dados: uma aplicação do algoritmo EM para o Reclink. Cad Saúde Colet (Rio J.) 2006; 14:225-32.
8. Winkler WE. Overview of record linkage and current research directions. Washington DC: Statistical Research Division, U.S. Census Bureau; 2006. (Research Report Series, 2006-2).
9. Teixeira CLS, Bloch KV, Klein CH, Coeli CM. Método de relacionamento de bancos de dados do Sistema de Informações sobre Mortalidade (SIM) e das autorizações de internação hospitalar (BDAIH) no Sistema Único de Saúde (SUS), na investigação de óbitos de causa mal definida no Estado do Rio de Janeiro. Epidemiol Serv Saúde 2006; 15:47-58.

Submetido em 02/Dez/2010
Versão final reapresentada em 01/Jun/2011
Aprovado em 22/Jun/2011