Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida

Accuracy of the probabilistic record linkage methodology to ascertain deaths in survival studies

Evandro Silva Freire Coutinho [1,2]
Cláudia Medina Coeli [2,3]

[1] *Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.*
[2] *Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.*
[3] *Escola Politécnica de Saúde Joaquim Venâncio, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.*

Correspondência
E. S. F. Coutinho
Departamento de Epidemiologia e Métodos Quantitativos em Saúde, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz.
Rua Leopoldo Bulhões 1480, Rio de Janeiro, RJ 21041-210, Brasil.
esfcoutinho@ensp.fiocruz.br

Abstract

Probabilistic record linkage methodology has been increasingly used to ascertain outcomes in cohort studies. However, only a few studies have evaluated its accuracy. The aim of this study was to evaluate the accuracy of probabilistic record linkage methodology to ascertain deaths in a cohort of 250 elderly people hospitalized for fractures caused by falls. The vital status of cohort members was determined through household visits conducted at one and six months after hospital admission. Record linkage was used to link records from the cohort to the Rio de Janeiro State mortality database. Deaths identified by the latter process were compared to the results obtained by active follow-up (gold standard). The vital status of 20 cohort members could not be determined by active follow-up. Among the remaining 230, 60 died within 180 days after hospital admission. Record linkage had a sensitivity of 85.5%, specificity of 99.4%, positive predictive value of 98.1%, and negative predictive value of 94.9%. In conclusion, probabilistic record linkage appears to be a useful tool for identifying deaths in cohort studies conducted in Brazil.

Records; Medical Record Linkage; Data Analysis

Introdução

O relacionamento de bancos de dados vem crescentemente sendo empregado para o seguimento passivo em estudos de coorte. Nessa forma de seguimento são utilizados bancos de dados epidemiológicos e administrativos para a identificação de desfechos, tais como óbitos, hospitalizações e casos incidentes de câncer [1].

Relacionar registros em diferentes bases de dados é uma tarefa trivial nos casos onde os registros de cada base incluam um campo comum, pouco sujeito a erros, que permita a identificação de cada ocorrência de forma unívoca (ex.: Cadastro de Pessoas Físicas – CPF). Na ausência desse identificador, o método probabilístico representa uma alternativa. Esse método é baseado, então, na utilização de campos comuns (ex.: nome, data de nascimento) presentes nas bases a serem relacionadas. Esses campos são utilizados conjuntamente para o cálculo de um escore, que traduz o grau de concordância entre os registros de cada par formado. Adicionalmente, deve ser definida uma regra (valores de escores limiares) para a classificação dos pares em verdadeiros, falsos ou duvidosos, devendo estes últimos serem submetidos à revisão manual para a classificação da situação final [2,3].

Sendo probabilístico, o processo é sujeito a erros, i.e., alguns pares podem ser classificados como verdadeiros quando na realidade os re-

gistros referem-se a indivíduos diferentes (erros homônimos ou falsos-positivos), enquanto outros podem não ser identificados como pares verdadeiros (erros sinônimos ou falsos-negativos). Os erros homônimos são mais freqüentes quando se utilizam poucos campos e os mesmos apresentam baixo poder discriminatório. Já os erros sinônimos derivam de falhas de registro (informação ou digitação) ou de mudanças nos identificadores pessoais (ex. mudança de sobrenome ou endereço) [4].

A avaliação da acurácia do método probabilístico pode ser realizada a partir da comparação dos resultados obtidos no processo de relacionamento com uma fonte de informação independente sobre a ocorrência dos desfechos de interesse (padrão-ouro) [5,6]. Apesar das potenciais vantagens desse tipo de abordagem, a mesma é pouco empregada em função da baixa disponibilidade de informações sobre os desfechos de interesse que possam ser consideradas como padrão-ouro.

O objetivo do presente estudo foi conhecer a acurácia da metodologia de relacionamento probabilístico de bases de dados para identificação de óbitos em uma coorte de idosos, empregando-se como padrão-ouro as informações colhidas por meio do seguimento ativo da coorte.

Metodologia

Uma coorte de 250 idosos (60 anos ou mais), internados por fratura decorrente de queda entre novembro de 2000 e fevereiro de 2002, com ou sem cirurgia, foi acompanhada originalmente para uma análise de sobrevida. Os idosos haviam sido internados em cinco hospitais públicos do Município do Rio de Janeiro, Brasil, e foram visitados 30 e 180 dias após a admissão. Nessas duas ocasiões, verificou-se se esses indivíduos estavam vivos ou não.

O banco de dados produzido para o estudo original, por meio de entrevista domiciliar, foi relacionado com o banco das declarações de óbitos do Estado do Rio de Janeiro de 2000 a 2003. As bases de dados de mortalidade foram obtidas junto ao Departamento de Dados Vitais, Secretaria de Estado de Saúde do Rio de Janeiro, para o projeto *Relacionamento de Grandes Bases de Dados em Saúde: Desenvolvimento e Aplicação da Metodologia do Relacionamento Probabilístico de Registros*, aprovado pelo Comitê de Ética em Pesquisa do Núcleo de Estudos de Saúde Coletiva, Universidade Federal do Rio de Janeiro.

O relacionamento foi realizado usando-se o programa RecLink II [7]. Inicialmente procedeu-se a padronização e a quebra em componentes dos campos nome e data de nascimento. Foi utilizada uma estratégia de relacionamento em cinco passos, empregando-se como chave de blocagem a combinação dos seguintes campos: *soundex* do primeiro nome, *soundex* do último nome, sexo e ano de nascimento. A blocagem permite que as bases de dados sejam logicamente divididas em blocos mutuamente exclusivos, segundo uma chave de indexação, sendo as comparações e cálculo de escores limitados aos registros pertencentes a um mesmo bloco [8].

Para o cálculo dos escores empregou-se os campos nome completo e data de nascimento, que foram comparados utilizando-se, respectivamente, algoritmos baseados na distância de Levenshtein, e na comparação caractere a caractere [7]. Foram utilizados valores de parâmetros de pareamento estimados a partir de uma amostra de treinamento [8] e apresentados em Camargo Jr. & Coeli [7]. Com essa configuração os valores de escore variaram entre -6,9 (discordância nos dois campos) até 10,7 (concordância total nos dois campos). Neste estudo não foi utilizado um valor de escore limiar para a classificação automática dos pares como verdadeiros, i.e., todos os pares com valores de escore superior a -3 foram revisados manualmente por um dos autores (C.M.C.) para a classificação final. Pares com valor de escore inferior a esse limiar foram considerados como falsos.

Utilizando-se a informação da entrevista como padrão-ouro, estimaram-se a sensibilidade, a especificidade e os valores preditivos para os dados obtidos por meio do relacionamento dos bancos. As medidas de acurácia foram apresentadas juntamente com os intervalos de confiança de 95%, calculados utilizando-se o método de Wilson [9].

Resultados

Dos 250 idosos que formavam a coorte de hospitalizados por fratura, houve 20 perdas decorrentes de recusa ou mudança de endereço. Dentre os 230 idosos para os quais foi possível obter informação nas visitas domiciliares, 60 (26,1%; IC95%: 20,5-32,3%) faleceram nos 180 dias que se seguiram à internação, sendo 37 nos primeiros 30 dias e 23 entre o 31º e 180º dia.

Na Tabela 1 encontram-se os dados de ocorrência de óbito na coorte, obtidos pela entrevista domiciliar (seguimento ativo – padrão-ouro) e pelo relacionamento dos bancos (seguimento passivo). O seguimento passivo não identificou nove óbitos (falsos-negativos) e classificou um idoso incorretamente como tendo falecido

Tabela 1

Acurácia na identificação de óbitos através de seguimento passivo baseado no relacionamento probabilístico de bases de dados em relação ao seguimento ativo (padrão-ouro).

Seguimento passivo	Seguimento ativo	
	Óbito	Vivo
Óbito	51	1
Vivo	9	169

Sensibilidade: 85% (IC95%: 73,9-91,9);
Especificidade: 99,4% (IC95%: 96,7-99,9);
Valor preditivo positivo: 98,1% (IC95%: 89,9-99,7);
Valor preditivo negativo: 94,9% (IC95%: 94,9-97,3).

(falso-positivo). A sensibilidade foi de 85,5%, enquanto a especificidade foi de 99,4%. Os valores preditivos positivo e negativo foram iguais a 98,1% e 94,9%, respectivamente.

Analisando-se os nove casos falsos-negativos, através da inspeção manual dos arquivos gerados pelo programa RecLink II e das bases de mortalidade, foi possível identificar três casos. Um deles apresentava um erro no primeiro nome fazendo com que o mesmo só pudesse ser recuperado no terceiro passo de blocagem (*soundex* do último nome + sexo). O escore final desse par foi alto, sendo o mesmo identificado como verdadeiro; entretanto, houve um erro de processamento no momento da atualização do arquivo de pares durante o passo, impedindo que o par fosse incorporado a esse arquivo. Para os dois casos restantes houve problema na identificação dos pares em função de erros no registro dos campos data de nascimento e nome. Em um deles faltava a informação sobre a data de nascimento e o nome era comum, dessa forma optou-se por não classificar o par como verdadeiro. Já no segundo caso, houve erro expressivo na base de mortalidade no registro tanto da data de nascimento como no campo nome completo.

Quanto ao registro falso positivo, o mesmo apresentava o nome igual nos dois registros e diferença apenas em um dígito da data de nascimento, levando à classificação incorreta como par.

Por fim, entre os vinte idosos perdidos no seguimento, foi possível identificar três óbitos por meio do relacionamento probabilístico de bases de dados.

Discussão

Os resultados deste estudo apontaram uma acurácia que pode ser considerada satisfatória, especialmente em relação à especificidade, do método do relacionamento probabilístico de registros para a identificação de óbitos através do seguimento passivo de uma coorte de pacientes idosos que foram internados por fratura decorrente de queda.

Erros homônimos e sinônimos influenciam de forma diferente as estimativas das medidas de associação entre variáveis explicativas e os desfechos de interesse em estudos de coorte. Nas situações onde a especificidade não varia segundo os níveis da variável de exposição de interesse (erros não-informativos), os resultados falsos-positivos tendem a enviesar a diferença e a razão de taxas em direção ao nulo. Já os resultados falsos-negativos, quando a sensibilidade é não diferencial em relação aos níveis da variável de exposição, tendem a enviesar em direção ao nulo a diferença, porém não a razão de taxas [1]. Adicionalmente, Brener et al. [4] chamam a atenção para o fato dos erros homônimos tenderem a aumentar com o crescimento do número de registros nas bases onde são buscados os desfechos. Sendo assim, recomenda-se que se opte por utilizar regras de classificação que minimizem a ocorrência de falsos-positivos [1,4,10].

Nosso resultado de especificidade foi comparável aos observados por outros autores que avaliaram a concordância do método do relacionamento probabilístico de registros em estudos realizados no Canadá [5] e Escócia [6], embora o resultado da sensibilidade tenha sido um pouco pior. Em função de contarmos com poucos campos para o relacionamento, optamos por usar uma estratégia que privilegiou a redução de falsos-positivos. Em pelo menos um dos casos o óbito não foi identificado em função de se tratar de nome comum. Chama atenção o fato de seis registros não identificados no processo de relacionamento automático não conseguirem ser encontrados em uma busca detalhada nas bases de óbito, o que poderia ser explicado por mudança de endereço, erros expressivos nos campos usados no processo de relacionamento e ausência de registros nas bases de óbito, já que as mesmas foram captadas antes da revisão e congelamento final das mesmas.

A inclusão de novos campos no processo de relacionamento automático poderia melhorar a performance do método. O bairro de residência e o nome da mãe seriam opções potenciais disponíveis nas bases de óbito. Com relação ao endereço, entretanto, a utilidade do mesmo de-

ve ser avaliada. Em um estudo onde dados de um inquérito sobre hospitalizações foram relacionados às bases de dados dos formulários de Autorização de Internação Hospitalar, o percentual de discordância no endereço nos pares verdadeiros foi elevado [11].

Uma vantagem da metodologia é a possibilidade da identificação da ocorrência dos desfechos de interesse mesmo entre os indivíduos perdidos de observação, como realizado no presente estudo.

Concluindo, o método do relacionamento probabilístico de registros representa uma ferramenta potencialmente útil para a identificação de desfechos em estudos de coorte desenvolvidos no Brasil.

Resumo

Poucos estudos avaliaram a acurácia do método do relacionamento probabilístico de registros para a identificação de desfechos em estudos de coorte. O objetivo deste estudo foi avaliar a acurácia do método para a identificação de óbitos em uma coorte de 250 idosos hospitalizados por fratura decorrente de queda. O estado vital dos membros da coorte foi determinado por meio de visitas domiciliares realizadas um e seis meses após a hospitalização. O método probabilístico foi usado para relacionar os registros da coorte (padrão-ouro) com aqueles da base de mortalidade do Estado do Rio de Janeiro, Brasil, visando à identificação de óbitos. O estado vital de vinte membros da coorte não pôde ser determinado. Entre os 230 restantes, 60 faleceram dentro do período de 180 dias após a admissão hospitalar. O método apresentou uma sensibilidade de 85,5%, uma especificidade de 99,4%, um valor preditivo positivo de 98,1% e um valor preditivo negativo de 94,9%. Concluindo, o relacionamento probabilístico de registros parece ser uma ferramenta útil para a identificação de óbitos em estudos de coorte realizados no Brasil.

Registros; Registro Médico Coordenado; Análise de Dados

Colaboradores

E. S. F. Coutinho idealizou e coordenou o estudo, analisou os dados e escreveu o artigo. C. M. Coeli realizou o relacionamento dos bancos de dados e escreveu o artigo.

Agradecimentos

Esta pesquisa recebeu apoio financeiro da Wellcome Trust e do Conselho Nacional de Desenvolvimento Científico e Tecnológico.

Referências

1. Howe G. Use of computerized record linkage in cohort studies. Epidemiol Rev 1998; 20:112-21.
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. Jaro MA. Probabilistic linkage of large public health. Stat Med 1995; 14:491-8.
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. 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; 80:54-7.
6. 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; 48:1441-52.
7. Camargo Jr. KR, Coeli CM. RecLink II: guia do usuário. http://planeta.terra.com.br/educacao/kencamargo/RecLinkII.html (acessado em 27/Jul/2005).
8. Coeli CM, Camargo Jr. KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol 2002; 5:185-96.
9. Altman DG. Diagnostic tests. In: Altman DG, Machin D, Bryant TN, Gardner MJ, editors. Statistics with confidence. 2nd Ed. London: BMJ Books; 2000. p. 105-19.
10. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol 2002; 31:1246-52.
11. Coeli CM, Blais R, Costa MC, Almeida LM. Probabilistic linkage in household survey on hospital care usage. Rev Saúde Pública 2003; 37:91-9.

Recebido em 24/Ago/2005
Versão final reapresentada em 26/Jun/2006
Aprovado em 18/Jul/2006