Revista de Saúde Pública
ISSN: 0034-8910
revsp@usp.br
Universidade de São Paulo
Brasil

Migowski, Arn; Brant Martins Chaves, Rogério; Medina Coeli, Cláudia; Pinho Ribeiro, Antonio Luiz; Rangel Tura, Bernardo; Caetano Kuschnir, Maria Cristina; Pereira Azevedo, Vitor Manuel; Brasil Floriano, Daniel; Moreira Magalhães, Carlos Alberto; Chagas Macedo Pinheiro, Márcia Cristina; de Aquino Xavier, Regina Maria

Acurácia do relacionamento probabilístico na avaliação da alta complexidade em cardiologia

Revista de Saúde Pública, vol. 45, núm. 2, abril, 2011, pp. 269-275
Universidade de São Paulo
São Paulo, Brasil

Rev Saúde Pública 2011;45(2):269-75

Arn Migowski[I]

Rogério Brant Martins Chaves[I]

Cláudia Medina Coeli[II]

Antonio Luiz Pinho Ribeiro[III]

Bernardo Rangel Tura[IV]

Maria Cristina Caetano Kuschnir[I]

Vitor Manuel Pereira Azevedo[I]

Daniel Brasil Floriano[IV]

Carlos Alberto Moreira Magalhães[I]

Márcia Cristina Chagas Macedo Pinheiro[I]

Regina Maria de Aquino Xavier[I]

Acurácia do relacionamento probabilístico na avaliação da alta complexidade em cardiologia

Accuracy of probabilistic record linkage in the assessment of high-complexity cardiology procedures

RESUMO

OBJETIVO: Avaliar a viabilidade de estratégia de relacionamento probabilístico de bases de dados na identificação de óbitos de pacientes submetidos a procedimentos de alta complexidade em cardiologia.

MÉTODOS: O custo de processamento foi estimado com base em 1.672 registros de pacientes submetidos à cirurgia de revascularização do miocárdio, relacionados com todos os registros de óbito no Brasil em 2005. A acurácia do relacionamento baseou-se em *linkage* probabilístico entre 99 registros de autorização de internação hospitalar de pacientes submetidos a cirurgias cardíacas em instituto de referência em cardiologia, com status vital conhecido, e todos os registros de óbito do estado do Rio de Janeiro em 2005. O *linkage* foi realizado em quatro etapas: padronização das bases, blocagem, pareamento e classificação dos pares. Utilizou-se a blocagem em cinco passos, com chaves de blocagem com combinação de variáveis como *soundex* do primeiro e último nome, sexo e ano de nascimento. As variáveis utilizadas no pareamento foram "nome completo", com a utilização da distância de Levenshtein, e "data de nascimento".

RESULTADOS: O segundo e o quinto passos de blocagem tiveram os maiores números de pares formados e os maiores tempos de processamento para o pareamento. O quarto passo demandou menor custo de processamento. No estudo de acurácia, após os cinco passos de blocagem, a sensibilidade do *linkage* foi de 90,6% e a especificidade foi de 100%.

CONCLUSÕES: A estratégia de relacionamento probabilístico utilizada apresenta boa acurácia e poderá ser utilizada em estudos sobre a efetividade dos procedimentos de alta complexidade e alto custo em cardiologia.

DESCRITORES: Procedimentos Cirúrgicos Cardiovasculares. Cardiopatias, mortalidade. Registros como Assunto. Registros de Mortalidade. Sistemas de Informação. Registro Médico Coordenado.

[I] Núcleo de Saúde Coletiva. Coordenação de Ensino e Pesquisa. Instituto Nacional de Cardiologia. Rio de Janeiro, Brasil

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

[III] Departamento de Clínica Médica. Faculdade de Medicina. Universidade Federal de Minas Gerais. Belo Horizonte, MG, Brasil

[IV] Núcleo de Bioestatística e Bioinformática. Coordenação de Ensino e Pesquisa. Instituto Nacional de Cardiologia. Rio de Janeiro, RJ, Brasil

Correspondência | Correspondence:
Arn Migowski
Núcleo de Saúde Coletiva
R. das Laranjeiras, 374, 5º andar
Laranjeiras
22240-006 Rio de Janeiro, RJ, Brasil
E-mail: arnmigowski@yahoo.com.br

Recebido: 17/12/2009
Aprovado: 25/8/2010

Artigo disponível em português e inglês em: www.scielo.br/rsp

ABSTRACT

OBJECTIVE: To evaluate the viability of a probabilistic record linkage strategy to identify patients who underwent complex cardiology procedures among the total deceased population.

METHODS: The processing cost was estimated based on 1,672 records of patients undergoing coronary artery bypass grafting that were compared with all death records in Brazil in 2005. The accuracy of the linkage strategy was based on the probabilistic linkage of 99 hospital admissions records of patients, with known vital status, who underwent cardiac surgery at a single cardiology institute, with the death records of the state of Rio de Janeiro, Southeastern Brazil, in 2005. Linkage was conducted in four stages: standardizing the databases, blocking, matching, and rating peers. Blocking in five steps was used, with blocking keys formed by a combination of variables such as soundex codes for the first and last names, sex, and year of birth. The variables used for matching were "full name" with the use of Levenshtein distance and "birth date".

RESULTS: The second and fifth blocking steps resulted in the largest number of formed pairs and the largest processing times for the matching. The fourth step required a lower processing cost. In the accuracy study, after five blocking steps, the sensitivity of the linkage was 90.6%, and the specificity was 100%.

CONCLUSIONS: The probabilistic strategy used has high accuracy and can be used in studies of the effectiveness of high-complexity, high-cost cardiology procedures.

DESCRIPTORS: Cardiovascular Surgical Procedures. Heart disease, mortality. Records as Topic. Mortality registries. Information Systems. Medical Record Linkage.

INTRODUÇÃO

Em função dos altos custos, a incorporação crítica de novas tecnologias na área de saúde, bem como a avaliação de tecnologias utilizadas em grande escala, consiste em uma das áreas de maior interesse para os sistemas de saúde em todo o mundo. A utilização de tecnologias fora das condições nas quais se mostraram eficazes é outro campo de crescente interesse.[14]

O incremento ascendente das despesas na assistência cardiovascular de alta complexidade fez com que o gasto total do Sistema Único de Saúde (SUS) aumentasse de aproximadamente R$ 395 milhões em 2000 para cerca de R$ 736 milhões em 2005, um acréscimo de 86,2%, segundo dados do Datasus. Foram gastos quase R$ 52 milhões com o tratamento de 1.549 pacientes candidatos ao implante de marca-passo de alto custo (desfibrilador implantável e/ou marca-passo multissítio) em 2005. No entanto, não é conhecida a efetividade de médio e longo prazo das intervenções especializadas promovidas e financiadas pelo SUS na área de cardiologia.

O Sistema de Informações Hospitalares (SIH-SUS) é a principal fonte de informações sobre os procedimentos cardiovasculares de alta complexidade pagos pelo SUS. O SIH-SUS é usado para estudar aspectos relacionados ao acesso e à qualidade do atendimento de alta complexidade em cardiologia.[a,b]

Apesar das limitações desse banco de dados, primariamente administrativo, a confiabilidade de dados de autorização de internação hospitalar (AIH) sobre o diagnóstico de doenças cardiovasculares é considerada satisfatória.[7,10]

Limitações do SIH quanto à informação sobre data e causa do óbito são superadas com a complementação dos dados pelos do Sistema de Informações de Mortalidade (SIM), tanto na identificação de óbitos após a alta hospitalar e em períodos de seguimento mais longos quanto na melhoria nos dados sobre a causa básica da morte.[11]

[a] Noronha JC. Utilização de indicadores de resultados para a avaliação da qualidade em hospitais de agudos: mortalidade hospitalar após cirurgia de revascularização do miocárdio em hospitais brasileiros [tese de doutorado]. Rio de Janeiro: Instituto de Medicina Social da UERJ; 2001.

[b] Lyra TM. O desafio da eqüidade no SUS: o uso do sistema de informações hospitalares na avaliação da distribuição da atenção cardiológica de alta complexidade, Brasil 1993-1999 [dissertação de mestrado]. Recife: Centro de Pesquisas Aggeu Magalhães da Fiocruz; 2001.

Em virtude da inexistência de um campo identificador unívoco, capaz de unir de forma determinística os registros do SIH e SIM, não é aproveitada a complementaridade dos dois sistemas. Para minimizar esse tipo de problema, procedimentos de relacionamento de bases de dados são desenvolvidos, sobretudo com uso de métodos probabilísticos.[2,13]

Todavia, poucos estudos avaliaram a acurácia de estratégias de relacionamento (*linkage)* probabilístico empregando bases de dados nacionais e nenhum foi realizado para avaliação do *linkage* entre o SIH-SUS e o SIM no contexto de procedimentos de alta complexidade em cardiologia.[15] O custo de processamento do relacionamento probabilístico, apesar de seu impacto na factibilidade do método, também é pouco estudado.

O objetivo deste estudo foi avaliar a viabilidade de estratégia de relacionamento probabilístico de bases de dados na identificação de óbitos de pacientes submetidos a procedimentos de alta complexidade em cardiologia.

MÉTODOS

As fontes de dados utilizadas foram a AIH do SIH-SUS do ano de 2005, como base de referência, e a base de declarações de óbito do SIM do mesmo ano, como base de comparação.

Foram realizados dois estudos. Em um deles, foram avaliados o custo de processamento e a adequação da estratégia de blocagem. O outro avaliou a acurácia do relacionamento probabilístico.

Para avaliar o custo de processamento e a eficácia na identificação de pares verdadeiros em cada passo de blocagem, foram selecionados 1.672 pacientes submetidos à cirurgia de revascularização do miocárdio (CRM) com AIH paga pelo SUS em todo o Brasil em 2005 e que evoluíram para óbito durante a internação.

A CRM foi escolhida por ser relativamente freqüente. Apenas óbitos hospitalares foram utilizados, uma vez que não havia padrão-ouro para determinar o *status* vital do grupo de pacientes após a internação. Foram consideradas as CRM isoladas e as associadas a outros procedimentos, como troca valvar, infartectomia ou aneurismectomia. O arquivo de declarações de óbito utilizado possuía 1.006.827 registros, correspondentes a todos os óbitos no Brasil no ano de 2005.

O número de pares possíveis com a combinação dessas duas bases de dados seria equivalente ao produto entre o número de registros das duas bases (1.683.414.744 pares). Com a blocagem, as comparações ficam limitadas a registros do mesmo bloco, ou seja, àqueles que possuem em comum valores de todas as variáveis contidas em cada chave de blocagem, o que diminui o total de pares formados a cada passo e aumenta a probabilidade de formação de pares verdadeiros.[5] Por diminuir o número de comparações feitas pelo computador, a blocagem diminui o uso da memória e do processador, o que resulta em menor custo de processamento e torna o processo mais rápido e menos exigente em termos de *hardware*. Um número muito elevado de pares poderia inviabilizar sua classificação manual.

Para avaliar a acurácia do relacionamento probabilístico, foram selecionados 452 pacientes consecutivos do banco de dados de cirurgias cardíacas do instituto em 2005. Foi realizado um *linkage* determinístico com o banco de AIHs apresentadas pelo instituto no período e foram localizados 353 registros. Para 128, não foi possível identificar o número de AIH relativo ao procedimento de interesse, uma vez que possuíam mais de uma AIH, restando 225 pacientes. Foram realizadas ligações telefônicas, com o intuito de conhecer o *status* vital dos 225 pacientes em março/abril de 2008, a data e o local do óbito, caso ocorrido.

Do total, 102 (45%) foram localizados na busca ativa, a despeito das diversas tentativas realizadas. Desses, oito haviam evoluído para óbito no período (dois em casa e o restante no próprio hospital) e 94 estavam vivos. Um *linkage* determinístico foi realizado para localizar as AIHs. O número da AIH foi utilizado para reunir os dados da busca ativa aos do SIH de 2005, fornecido pelo Datasus.

Foram localizados 45 dos 102 pacientes (44%) no banco do Datasus: 44 vivos em 2008 e um óbito durante a internação. A esses, juntaram-se 55 pacientes submetidos a cirurgias cardíacas com óbito hospitalar conhecido por meio dos arquivos de AIH, totalizando 99 pacientes.

Os arquivos de AIH, fornecidos pelo Datasus por mês e unidade federativa, foram reunidos e novo banco de dados foi criado no MySQL.[c] A seleção das variáveis dos arquivos do SIM e do SIH-SUS e dos conjuntos de registros que foram utilizados em cada passo do processo de *linkage*, bem como a homogeneização do tamanho e codificação das variáveis das duas bases de dados foram realizadas por meio do *software* Delphi.[d]

O *linkage* probabilístico foi realizado empregando-se a terceira versão do *software* RecLink.[2,3] As bases a serem relacionadas foram padronizadas para alcançar equivalência de grafia, formatos e conteúdos. Foi utilizada estratégia de blocagem em cinco passos, a partir da combinação dos seguintes campos: *soundex* do primeiro nome (modificado), *soundex* do último nome (modificado), sexo e ano de nascimento.[5]

[c] MySQL. The world's most popular open source database [Internet]. Sweden: Oracle; 2009[citado 2009 ago 6]. Disponível em: http://www.mysql.com

[d] Delphi 7: Database Desktop Tool [CD-ROM]. Version 7.0. Austin: Borland International Inc; 1992.

Em seguida, foi realizado o pareamento ou *matching,* em que são formados pares (um registro do SIH com um registro do SIM) a partir da comparação de variáveis previamente selecionadas. Foram utilizadas as variáveis "nome completo" e "data do nascimento". Escores foram gerados para cada par, com base na similaridade entre os valores das variáveis selecionadas. Para comparação da variável "nome completo" entre pares, foi utilizada a distância de Levenshtein e para a variável "data do nascimento" utilizou-se um algoritmo para caractere.[2] Os parâmetros para a construção dos fatores de ponderação (concordância ou discordância) foram estimados com base no algoritmo *expectation-maximization* pelo próprio *software* RecLink.[12]

Revisão e classificação manual dos pares foram realizadas por dois pesquisadores. Os pares foram classificados como "verdadeiros", "duvidosos" ou "falsos". A revisão exaustiva de todos os pares foi realizada nas faixas com maior escore, na faixa de escore formada por pares com nomes idênticos e pontuação intermediária e na faixa de escore com data de nascimento idêntica e pontuação baixa. Nas faixas com escores muito baixos, os pares foram considerados automaticamente como falsos pares.

Um conjunto prévio de critérios foi estabelecido para classificação dos pares como verdadeiros ou falsos. Os critérios utilizados envolveram a raridade dos nomes e sobrenomes, o grau de concordância dos nomes, das datas de nascimento e do município de residência.

Os registros dos pares, considerados verdadeiros no primeiro passo, foram retirados das duas bases de dados (Tabela 1) e o segundo passo de blocagem foi realizado, seguido por novo pareamento e nova etapa de classificação dos pares. Os passos seguintes foram realizados sucessivamente até o quinto.

O computador utilizado apresentava as seguintes características: processador de quatro núcleos de tecnologia de 45 nm e *clock* de 2,83 GHz, 8 GB de memória, 1.333 MHz em *Dual Channel* e HD SATA 2 de 300GB e velocidade de 10.000 rpm. Utilizando o *software performance test trial,* da *PassMark software Inc.,* obteve-se pontuação de 1.306. Isso significa que o computador utilizado realiza suas operações 1.306 vezes mais rápido do que o ATX 286, utilizado como base de comparação, possibilitando o *benchmarking* com outros computadores.

Os indicadores utilizados na avaliação do custo de processamento foram o número de pares formados em cada passo e o tempo gasto em cada etapa do processo. Além de refletir o custo de processamento, o número de pares formados em cada passo foi indicador de eficiência da etapa de blocagem.

Classificamos o total de pares formados em cada etapa como "pares verdadeiros", "zona cinza" ou "pares falsos". A "zona cinza" refere-se às faixas de escore em que não foi possível classificar todos os pares nelas contidos como verdadeiros ou falsos, em virtude de sua heterogeneidade interna. Em função disso, foi necessária a classificação manual de cada par nessas faixas. Portanto, o número de pares na zona cinza (Tabela 2) foi indicador da carga de trabalho na revisão manual dos pares duvidosos.

Para a avaliação da acurácia do método, foram utilizadas a sensibilidade e a especificidade, valor preditivo positivo e negativo, considerando como padrão-ouro a informação sobre o *status* vital de cada paciente obtida na busca ativa ou na mortalidade hospitalar nas AIHs. O padrão-ouro é definido como uma fonte externa da "verdade" com relação à situação de doença – neste caso o *status* vital – em cada indivíduo da população de estudo.[8] Na etapa de classificação manual dos pares, os pesquisadores estavam cegos para o *status* vital dado pelo padrão-ouro.

Duas AIHs com nomes inespecíficos iniciados por "RN de" seguidos pelo nome da mãe foram excluídas do estudo de acurácia pela impossibilidade de localizá-los no banco de mortalidade. Intervalos com 95% de confiança foram calculados pelo método de Wilson[1] com o pacote PropCIs, versão 0.1-6 para o *software* R.

A pesquisa foi aprovada previamente por comitê de ética da Universidade Federal de Minas Gerais (em 4/5/2009, protocolo nº 0084.0.203.000-09).

RESULTADOS

Os passos 2 e 5 foram os que apresentaram maior tempo de execução do pareamento por apresentarem chaves de blocagem menos específicas (Tabela 1). Isso foi confirmado pelo maior número de pares formados nesses dois passos (Tabela 2). Os passos 1 e 2, em que estava concentrado o maior número de pares verdadeiros, demandaram maior tempo de processamento na formação dos arquivos oriundos do *linkage* (Tabela 1).

A estratégia de seleção de pontos de corte e classificação automática de pares, denominada "passo 0", consumiu 42 minutos após o primeiro passo de blocagem, enquanto a estratégia alternativa de revisão manual dos pares levou duas horas para ser concluída (Tabela 1). A estratégia de revisão manual resultou em sensibilidade de 80%, enquanto a estratégia de classificação "automática" atingiu 72% (Tabela 2).

Na classificação manual dos pares, a localização dos pares verdadeiros não foi homogênea nas dezenas de faixas de escore. Foram identificados três tipos de faixas de interesse: (A) as primeiras faixas de escore, em que a pontuação alta resultou da grande semelhança das variáveis de pareamento (nome completo e data de nascimento); (B) uma faixa com escore intermediário, nomes idênticos e datas de nascimento bastante diferentes; (C) uma faixa com escore baixo, datas

Tabela 1. Tempo de processamento no *linkage* entre as bases de sistemas de informação, de acordo com o passo de blocagem.

Etapa	Tempo de Processamento (min)					
	Passo 0[a]	Passo 1	Passo 2	Passo 3	Passo 4	Passo 5
Estimação dos parâmetros de pareamento	12	12	0	0	0	0
Execução do pareamento (formação de pares com escores)	2	2	16	8	0,27	9
Seleção manual de pontos de corte p/ classificação automática de pares	30	0	0	0	0	0
Classificação automática dos pares	12	0	0	0	0	0
Classificação manual dos pares	0	120	105	20	25	15
Criação de arquivo de pares verdadeiros e novos arquivos SIH e SIM	30	32	6	4	3	5
Total	88	166	127	32	28	29

[a] O passo 0 representa apenas uma abordagem alternativa para o passo 1.
SIH: Sistema de Informações Hospitalares
SIM: Sistema de Informações de Mortalidade

de nascimento idênticas e nomes bastante diferentes (pela distância de Levenshtein). Neste último tipo, a diferença entre nomes completos pode corresponder à abreviação ou à omissão de nomes intermediários, o que tem grande impacto na distância de Levenshtein, mas é facilmente identificável na classificação manual.

Dos 1.411 pares verdadeiros encontrados no estudo de custo de processamento (Tabela 2), 97,9% foram localizados nas faixas com as características descritas em A, 2% em faixas semelhantes às descritas em C e 0,1% nas faixas do tipo descrito em B.

Em ambos os estudos, o primeiro passo de blocagem foi responsável pela localização da maioria dos pares verdadeiros (95% do total de pares verdadeiros no estudo de custo de processamento e 96% no estudo de acurácia) (Tabela 3).

No estudo de acurácia, a sensibilidade do método foi de 90,6% (Tabela 4). Cinco pacientes, que sabidamente haviam evoluído para óbito no período, não foram localizados pelo relacionamento probabilístico. Nenhum paciente sabidamente vivo ao final do seguimento foi classificado como óbito (falso-positivo) no estudo de acurácia, resultando em especificidade de 100% (Tabela 4).

DISCUSSÃO

O presente estudo mostrou bom resultado na acurácia da estratégia de *linkage* probabilístico utilizada e custo de processamento aceitável. O número de óbitos foi pouco subestimado com o uso do método (cinco falso-negativos) e nenhum paciente vivo ao final do seguimento foi classificado como morto (falso-positivos), mesmo na ausência de variáveis que aumentariam a especificidade da revisão manual, como "nome da mãe" e "município de nascimento". Como o evento "óbito" é relativamente raro nos procedimentos estudados, pequenos erros na especificidade teriam grande impacto na qualidade do linkage.

O baixo tempo de processamento está provavelmente associado à eficiência da estratégia de blocagem em múltiplos passos e do uso de chaves de blocagem com múltiplas variáveis, além da configuração adequada do computador utilizado. O número de procedimentos usados no estudo de custo de processamento é aplicável para a avaliação anual da maioria dos procedimentos cardiovasculares de alta complexidade, exceto angioplastia coronariana e cirurgia de revascularização do miocárdio, que são mais freqüentes.

Tabela 2. Resultados da aplicação da estratégia de blocagem em múltiplos passos.

Passos	Número de pares formados no passo	Número de pares na zona cinza	Número de pares verdadeiros no passo	Sensibilidade do *linkage* (%)
0 (igual ao passo 1, porém sem inspeção manual da zona cinza)	38.712	14.043	1.205	72
1: *soundex* do primeiro nome (modificado) + *soundex* do último nome (modificado) + sexo	38.712	14.043	1.340	80
2: *soundex* do primeiro nome (modificado) + sexo	146.070	3.812	32	82
3: *soundex* do último nome (modificado) + sexo	91.324	4.453	11	83
4: *soundex* do primeiro nome (modificado) + *soundex* do último nome (modificado)	9.024	3.570	26	84
5: ano de nascimento + sexo	695.253	69	2	84

Tabela 3. Resultados da aplicação da estratégia de blocagem em múltiplos passos no estudo de acurácia.

Passos	Número de pares verdadeiros no passo	Sensibilidade do *linkage* (%)	Especificidade do *linkage* (%)
1: *soundex* do primeiro nome (modificado) + *soundex* do último nome (modificado) + sexo	46	86,8	100
2: *soundex* do primeiro nome (modificado) + sexo	1	88,7	100
3: *soundex* do último nome (modificado) + sexo	0	88,7	100
4: *soundex* do primeiro nome (modificado) + *soundex* do último nome (modificado)	1	90,6	100
5: ano de nascimento + sexo	0	90,6	100

O quarto passo apresentou tempo de processamento consideravelmente menor do que os outros e resultados importantes, uma vez que possibilitou a captação de pares verdadeiros onde havia erro de digitação na variável sexo, por ser a única chave de blocagem sem essa variável.

O quinto passo de blocagem praticamente não aumentou a sensibilidade do método e possibilitou a localização de dois pares verdadeiros, que, apesar de terem aparecido na "zona cinza" de outras etapas, não tinham sido identificados. Embora seja uma das etapas com maior exigência de processamento pelo grande número de pares formados, a facilidade de identificar falsos pares (ex.: registros com dados faltantes) torna a inspeção manual menos trabalhosa. Essa etapa pode ser utilizada quando as bases usadas não forem muito grandes e houver necessidade de aumentar a sensibilidade do *linkage*.

A classificação manual dos pares foi a etapa que consumiu maior tempo. Essa etapa teve grande impacto sobre a sensibilidade do método. Além de responsável por aumento de 8% na sensibilidade no primeiro passo, foi o único método utilizado para a determinação do *status* dos pares nos passos 2, 3, 4 e 5, em função da impossibilidade de se estabelecerem pontos de corte para pares verdadeiros com bom poder de discriminação nos referidos passos. A inclusão de elevado número de óbitos hospitalares (55 pacientes) poderia aumentar a proporção de óbitos encontrados com o *linkage* devido à possível melhor qualidade do preenchimento da Declaração de Óbito, o que aumentaria a sensibilidade.

A necessidade de classificação manual dos pares em *linkages* envolvendo maior número de registros ou em que grande número de etapas seja requerido poderia inviabilizar o processo. Segundo nossos resultados, uma opção seria realizar a inspeção seletiva das faixas de escore de maior interesse, com característica de fácil identificação e responsáveis por 100% dos pares localizados. Outra estratégia aceitável, quando a sensibilidade do relacionamento probabilístico não for crítica, seria a realização apenas do primeiro passo de blocagem, tendo em vista que este foi responsável pela localização de 95% dos pares verdadeiros em nossos estudos. O papel dominante do primeiro passo de blocagem também foi constatado por outros pesquisadores.[5] Pesquisadores têm investido em técnicas automáticas para a diminuição do número de pares para revisão manual.[9]

Tabela 4. Acurácia na identificação de óbitos pelo relacionamento probabilístico.

Seguimento passivo (*linkage*)	Seguimento ativo (padrão-ouro)	
	Óbito	Vivo
Óbito	48	0
Vivo	5	44

Sensibilidade = 90,6% (IC95%: 79,7;95,9)
Especificidade = 100 % (IC95%: 92,0;100)
Valor preditivo positivo = 100% (IC95%: 92,6;100)
Valor preditivo negativo = 89,8% (IC95%: 78,2;95,6)

Os resultados de sensibilidade e especificidade do método foram semelhantes aos de outros estudos que envolveram relacionamento probabilístico de bases de dados conduzidos no Brasil, Nova Zelândia, Estados Unidos e Reino Unido.[15] Os resultados de acurácia encontrados foram semelhantes aos descritos por outros pesquisadores no *linkage* de dados primários com os registros do SIM.[6] Os valores preditivos, no entanto, tendem a ser diferentes em situações reais, uma vez que a prevalência do desfecho óbito foi artificialmente aumentada no estudo.

A questão da representatividade da base de AIHs do Datasus é suscitada pela dificuldade de localização de registros de AIH no teste de acurácia e deve ser mais bem investigada. Dificuldades na identificação de internações hospitalares na base do SIH foram descritas por outros pesquisadores.[4]

A estratégia de relacionamento probabilístico utilizada apresentou acurácia satisfatória e poderá ser empregada em estudos sobre a efetividade dos procedimentos de alta complexidade e alto custo em cardiologia. O quinto passo de blocagem resultou em custo de processamento excessivo e desprezível contribuição para a sensibilidade do método. A revisão manual dos pares é o ponto crítico do processo em termos de tempo, o que indica a necessidade de maior sistematização dessa etapa e de estudos que aperfeiçoem o método de cálculo de escores e aumentem seu poder discriminante.

AGRADECIMENTOS

À Dra. Regina Elizabeth Müller, do Instituto Nacional de Cardiologia, pelas sugestões ao texto, e a Maria Lúcia Zurita Monteiro, do Instituto Nacional de Cardiologia, pela participação no processo de busca ativa e sugestões.

REFERÊNCIAS

1. Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with Confidence. 2. ed. Bristol: British Medical Journal Books; 2000.
2. 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.
3. Camargo Jr KR, Coeli CM. Reclink 3: nova versão do programa que implementa a técnica de associação probabilística de registros (probabilistic record linkage). *Cad Saude Coletiva (Rio J)*. 2006;14(2):399-404.
4. Coeli CM, Blais R, Costa MCE, Almeida LM. Probabilistic linkage in household survey on hospital care usage. *Rev Saude Publica*. 2003;37(1):91-9. DOI:10.1590/S0034-89102003000100014
5. Coeli CM, Camargo Jr KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. *Rev Bras Epidemiol*. 2002;5(2):185-96. DOI:10.1590/S1415-790X2002000200006
6. Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida. *Cad Saude Publica*. 2006;22(10):2249-52. DOI:10.1590/S0102-311X2006001000031
7. Escosteguy CC, Portela MC, Medronho RA, Vasconcellos MTL. O Sistema de Informações Hospitalares e a assistência ao infarto agudo do miocárdio. *Rev Saude Publica*. 2002;36(4):491-9. DOI:10.1590/S0034-89102002000400016
8. Gordis L. Epidemiology. 4. ed. Philadelphia: Saunders Elsevier; 2009.
9. Machado CJ, Hill K. Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem. *Cad Saude Publica*. 2004;20(4):915-25. DOI:10.1590/S0102-311X2004000400005
10. Mathias TAF, Soboll MLMS. Confiabilidade de diagnósticos nos formulários de autorização de internação hospitalar. *Rev Saude Publica*. 1998;32(6):526-32. DOI:10.1590/S0034-89101998000600005
11. Melo ECP, Travassos C, Carvalho MS. Qualidade dos dados sobre óbitos por infarto agudo do miocárdio, Rio de Janeiro. *Rev Saude Publica*. 2004;38(3):385-91. DOI:10.1590/S0034-89102004000300008
12. Junger WL. Estimação de parâmetros em relacionamento probabilístico de bancos de dados: uma aplicação do algoritmo EM para o Reclink. *Cad Saude Colet (Rio J)*. 2006;14(2):225-32.
13. Silva JPL, Travassos C, Vasconcellos MM, Campos LM. Revisão sistemática sobre encadeamento ou linkage de bases de dados secundários para uso em pesquisa em saúde no Brasil. *Cad Saude Colet (Rio J)*. 2006;14(2):197-224.
14. Silva LK. Avaliação tecnológica e análise custo-efetividade em saúde: a incorporação de tecnologias e a produção de diretrizes clínicas para o SUS. *Cienc Saude Coletiva*. 2003;8(2):501-20. DOI:10.1590/S1413-81232003000200014
15. Silveira DP, Artmann E. Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática. *Rev Saude Publica*. 2009;43(5):875-82. DOI:10.1590/S0034-89102009005000060

Pesquisa financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Departamento de Ciência e Tecnologia do Ministério da Saúde (DECIT) (Nº PROCESSO: 551402/2007-5).
Trabalho apresentado no IX Congresso Brasileiro de Saúde Coletiva, realizado em Recife, PE, em 2009.
Os autores declaram não haver conflitos de interesse.