CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

Avaliação de método para classificação automatizada de pares em relacionamentos probabilísticos de bancos de dados

Assessment of a method for automatic match classification in probabilistic data linkage

Evaluación del método para la clasificación automatizada de pares en relaciones probabilísticas de bancos de datos

Daniela de Almeida Pereira Duarte [1,2]
Camila Soares Lima Corrêa [1]
Vívian Assis Fayer [1]
Mário Círio Nogueira [1]
Maria Teresa Bustamante-Teixeira [1]

doi: 10.1590/0102-311X00066419

Resumo

O objetivo foi testar e avaliar a acurácia de um método para a seleção de escore em relacionamento probabilístico de banco de dados, de forma a viabilizar a automatização da identificação de pares verdadeiros dispensando a etapa de inspeção manual. Estudo de acurácia utilizando dados do Sistema de Informação do Câncer de Mama (SISMAMA) de Minas Gerais, Brasil, de 2009 e 2010. Após o processo de limpeza e padronização, foi realizado o relacionamento probabilístico dos bancos 2009 e 2010 utilizando 16 passos, sendo que cada passo foi inspecionado manualmente para se obter um padrão-ouro. Posteriormente, selecionaram-se amostras que foram inspecionadas e avaliadas para calcular a acurácia do método de seleção dos pares verdadeiros. Todos os passos e amostras com 200 e 300 pares apresentaram alta sensibilidade (recall) *> 0,97, alto valor preditivo positivo* (precision) *> 0,95 e altas acurácia (> 0,97), medida F (> 0,96) e área sob a curva* precision-recall *(> 0,98). A amostra com 100 pares evidenciou altos valores para essas medidas, porém com escores mais baixos. Dos 16 passos avaliados, o uso de apenas três de forma combinada foi suficiente para identificar 99,24% dos pares verdadeiros no banco total. O método proposto permite automatizar o relacionamento das bases de dados, mantendo a acurácia do método. Facilita a utilização de relacionamento probabilístico no âmbito dos serviços de saúde, especialmente para a vigilância e gestão em saúde.*

Sistemas de Informação em Saúde; Integração de Sistemas; Confiabilidade dos Dados

Correspondência
D. A. P. Duarte
Rua Guanabara 578, Ponte Nova, MG 35430-098, Brasil.
danalmeidap@yahoo.com.br

[1] Universidade Federal de Juiz de Fora, Juiz de Fora, Brasil.
[2] Divisão de Saúde, Universidade Federal de Viçosa, Viçosa, Brasil.

Introdução

Os sistemas de informação em saúde (SIS) existentes no Brasil são numerosos e diversificados e a maioria foi criada para atender as necessidades de gerenciamento financeiro e/ou clínico. Avanços no campo da Saúde Coletiva, nos conceitos de saúde e qualidade de vida da população exigem cada vez mais conhecimento dos profissionais acerca da realidade social e de saúde das pessoas, com a finalidade de planejamento e priorização de suas ações. Dessa maneira, os sistemas de informação em saúde passaram a ser grandes aliados dos gestores e profissionais da saúde.

Cada setor da sociedade produz informações específicas para sua área de atuação. Para o setor saúde, a vinculação de dados de indivíduos de uma determinada população utilizando diferentes SIS é importante para se estabelecer e avaliar uma linha de cuidados, já que a saúde é resultante das condições sociais, políticas e econômicas do meio onde cada indivíduo está inserido.

O relacionamento de bases de dados é definido como a estratégia que reúne duas ou mais informações, registradas em diferentes fontes e que se referem a um mesmo indivíduo. Vincular essas fontes nem sempre é uma tarefa fácil e rápida, uma vez que algumas bases não apresentam um identificador único, como o Cartão Nacional de Saúde (CNS) do Sistema Único de Saúde (SUS) ou o Cadastro de Pessoas Físicas (CPF), que permitiriam a realização de um relacionamento determinístico. Além disso, há possibilidade de erros de registros, presença de homônimos, incompletude dos dados e mudanças de endereço. Dessa forma, uma solução é o relacionamento de bases por meio do método probabilístico, utilizando variáveis demográficas como nome, nome da mãe, data de nascimento e endereço, que pondera a vinculação destas informações, resultando na construção de escores de classificação [1].

Posteriormente é preciso estabelecer se as ligações estão ou não corretas. A classificação manual de pares usada nesse procedimento é considerada padrão-ouro. Porém, é uma etapa que demanda muito tempo, principalmente quando existe grande número de pares a ser inspecionados, situação que pode inviabilizar o processo [2]. Em um estudo sobre relacionamento de dados referentes ao setor de alta complexidade em cardiologia foram despendidas duas horas para conferir todos os pares formados, gerando a recomendação de que nos casos em que exista grande número de pares após o relacionamento seja realizada inspeção seletiva nas faixas de escore de maior interesse [3]. Outros estudos utilizaram escores para os limiares superior e inferior [4,5,6], indicando para a inspeção os pares com escores intermediários. Camargo Jr. & Coeli [7] sugeriram a definição do ponto de corte para o limiar inferior de forma arbitrária, determinando que registros com escore negativo deveriam ser considerados falsos pares e com escores positivos deveriam ser revistos. No entanto, essas práticas tornam-se infactíveis quando um grande número de pares atende ao critério de revisão. Uma terceira estratégia seria automatizar a categorização dos pares como verdadeiros ou falsos, dispensando a inspeção manual, entretanto, permaneceria a necessidade da definição de um valor de escore como ponto de corte [2].

Diante da necessidade de otimização do processo e de uma elevada acurácia dos resultados, é preciso que para cada relacionamento de dados seja estabelecido um escore específico que identifique os pares verdadeiros. Mas como definir esse valor? Estudos sugerem que o escore e a validade das estratégias de relacionamento sejam determinados baseando-se na comparação com um conjunto de dados referência cujo status de correspondência verdadeira seja conhecido, isto é, um padrão-ouro [1,8,9].

Esse padrão pode ser procedente de uma fonte de dados com identificadores completos ou de uma subamostra dos registros que foram revisados manualmente [9]. Revisão sistemática sobre a acurácia de relacionamento probabilístico de bases em saúde também sugere a geração de uma amostra e revisão manual dos pares desta, porém, não apresenta nenhum estudo em que este método foi aplicado [2]. Outro trabalho utilizou como padrão-ouro os bancos de pares verdadeiros, originados inicialmente no processo de relacionamento, que tinham como identificador único de cada indivíduo o CNS e/ou o CPF [4].

Para avaliar a qualidade e ilustrar o desempenho desses métodos geralmente são utilizados parâmetros como sensibilidade e especificidade e a curva ROC (*receiver operating characteristics*). Porém, quando os dados são desbalanceados/desequilibrados, como nos relacionamentos entre banco de dados, em que maior quantidade de pares verdadeiros negativos são formados em relação ao número de pares verdadeiros positivos, essas medidas têm se mostrado pouco precisas e informativas, sendo recomendado o uso da curva *precision-recall* (PRC) e da medida F. A curva PRC avalia a proporção

de pares verdadeiros positivos entre as previsões positivas, fornecendo uma visão mais precisa do desempenho do relacionamento [2,7,9,10,11].

Valor preditivo positivo (*precision*) é definido como a proporção de pares formados que estão corretos. Sensibilidade (*recall*) é a proporção de todos os pares verdadeiros que foram vinculados corretamente [12]. Tais medidas podem ser estimadas adotando-se como padrão-ouro a revisão manual [7,9,10].

Outro parâmetro utilizado para avaliar a qualidade do *linkage* é a medida F, uma média harmônica dos parâmetros *precision* e *recall*. Altos valores de *precision* e de *recall* levam a uma medida F elevada, que permite a definição de um escore de alto poder discriminatório na identificação de pares verdadeiros [13,14].

Este trabalho tem como objetivo testar e avaliar a acurácia de um método para a seleção de escore em relacionamento probabilístico de banco de dados, com base no qual seja possível identificar de forma automatizada todos os pares verdadeiros, dispensando a inspeção manual de todos os pares formados.

Metodologia

Estudo de acurácia de um método automatizado para a classificação dos pares formados baseando-se no relacionamento probabilístico, realizado com dados do Sistema de Informação do Câncer de Mama (SISMAMA) de Minas Gerais, Brasil. O banco de dados para o relacionamento foi construído valendo-se de exames alterados (BI-RADS 4, 5 e 6) do ano de 2009, e este foi comparado com o ano de 2010. Foram selecionados apenas os exames de mulheres residentes em Minas Gerais com 18 anos ou mais.

A padronização das bases de dados consistiu na remoção de acentos, cedilhas, espaços duplos, caracteres especiais tais como interrogação, exclamação e outros. Em seguida, foram excluídos os exames duplicados (mesmo nome, nome da mãe, data de nascimento e data de exame) e repetidos (mesmo nome, nome da mãe e data de nascimento) com a finalidade de manter apenas um registro de cada indivíduo.

O relacionamento probabilístico foi realizado em 16 passos e as chaves de blocagem e pareamento de cada um dos passos foram adaptadas de um estudo que usou os dados do SISMAMA [4] e seguiu as recomendações do manual do RecLink III [15] (Tabela 1). Diferentemente de outros trabalhos [3,4] nos quais a cada passo foram mantidos apenas os registros classificados como não par, no presente estudo todos os registros foram testados em todos os passos. Além disso, esses estudos utilizaram como padrão-ouro dados gerados no primeiro passo, cuja chave de blocagem usava identificador único (CPF e/ou CNS) [4] e revisão manual [3]. Neste trabalho, cada passo foi inspecionado manualmente com a finalidade de se estabelecer um padrão-ouro de ponto de corte para os pares verdadeiros. Os critérios para a definição de pares verdadeiros foram baseados em Girianelli et al. [16].

De cada passo foram extraídas três diferentes amostras aleatórias de pares, compostas por 100, 200 e 300 pares, respectivamente, para posterior revisão manual e comparação com o padrão-ouro. Para a definição do tamanho da amostra foi utilizada a função *power.roc.teste* do pacote pROC do programa R (http://www.r-project.org), que estimou que um tamanho amostral de cerca de 100 registros seria suficiente. Para fins de comparação foram testadas ainda amostras com 200 e 300 pares.

Para avaliar a acurácia do método em cada passo e nas amostras foram usadas a sensibilidade (*recall*) e o valor preditivo positivo (*precision*) por meio da curva *precision-recall* e da medida F, que possibilitaram a determinação do escore de maior poder discriminatório na identificação dos pares verdadeiros [11]. Tal análise utilizou os pacotes PRROC e ROCR do programa R (http://www.r-project.org).

Por fim, foi feita a combinação de todos os passos para encontrar o número total de pares verdadeiros. Procedeu-se, então, a combinação desses passos entre si para investigar quais deles em conjunto abrangem o maior número de pares verdadeiros. O programa R foi usado na limpeza dos dados, remoção das duplicidades e repetições, cálculo das amostras e construção da curva *precision-recall*, e o programa RecLink III [15] no relacionamento e também na remoção das duplicidades e repetições.

Este projeto foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de Juiz de Fora sob o número de parecer 1.431.916.

Resultados

Inicialmente os bancos 2009 e 2010 tinham 165.115 e 490.535 registros, respectivamente. Do número total de observações encontradas no banco referência (2009), 3.030 foram classificadas como BI-RADS 4, 5 e 6. Após a execução do processo de limpeza permaneceram nos bancos 2.764 (2009) e 445.531 (2010) registros.

No relacionamento foi constatado que o passo cuja variável de blocagem era ano de nascimento e as variáveis de pareamento eram nome, data de nascimento e nome da mãe (passo 16), formou o maior número de pares verdadeiros e o passo que utilizava o CNS como variável de blocagem e as mesmas várias de pareamento (passo 3) formou o menor número de pares verdadeiros (Tabela 1). Foi observado que o escore do relacionamento é influenciado pelo número de variáveis utilizadas no pareamento. Após a combinação de todos os passos chegou-se ao número total de 1.194 pares verdadeiros. Constatou-se que com apenas três passos é possível encontrar 99,24% dos pares verdadeiros e que com base em quatro passos poucos pares verdadeiros são acrescentados, no caso deste estudo um a três pares a cada passo incluído (Tabela 2).

No que se refere ao ponto de corte para a classificação de pares verdadeiros, a maioria dos passos tem escore acima de 6,5 e as amostras com 200 e 300 pares apresentam valores bem próximos a estes, variando entre 6,5 e 6,8. Houve pouca diferença nos escores das amostras com 200 e 300 pares (Tabela 3).

Todos os pares formados no passo 3, cuja variável de blocagem era CNS e de pareamento nome, data de nascimento e nome da mãe, eram verdadeiros e apenas três registros tinham baixos escores. A inexatidão encontrada em um dos pares estava no último nome e mês de nascimento; em outro o nome do pai foi registrado no local destinado para o registro do nome da mãe; e no terceiro registro a letra M do nome e do nome da mãe foi trocada por N e o último sobrenome era diferente, contudo, as demais variáveis eram iguais: data de nascimento, endereço, código do município e CNS.

Todos os passos e amostras com 200 e 300 pares apresentaram alta sensibilidade (*recall*) > 0,97, alto valor preditivo positivo (*precision*) > 0,95 e altas acurácia (> 0,97), medida F (> 0,96) e área sob a curva

Tabela 1

Passos, variáveis de blocagem e pareamento, escores, pares formados e pares verdadeiros do relacionamento dos bancos do Sistema de Informação do Câncer de Mama (SISMAMA), Minas Gerais, Brasil, 2009 e 2010.

Passos	Variáveis de blocagem	Variáveis de pareamento	Escore	Pares formados	Pares verdadeiros
1	PN + UN + AN	N + DN + M + MUN	-13 a 21	2.108	1.076
2	PN + UN + AN	N + DN	-6 a 10	6.067	1.069
3	CNS	N + DN + M	-10 a 17	216	216
4	PN + UN + AN + PM + UM	N + DN + M	-10 a 17	932	923
5	PN + UN + AN	N + DN + M	-10 a 17	1.373	1.069
6	PN + AN + PM + UM	N + DN + M	-10 a 17	1.067	983
7	PN + AN + UM	N + DN + M	-10 a 17	1.188	1.022
8	PN + AN + PM	N + DN + M	-10 a 17	1.223	1.075
9	PN + AN	N + DN + M	-10 a 17	2.620	1.142
10	PN + UN + PM + UM	N + DN + M	-10 a 17	991	950
11	PN + UN + PM	N + DN + M	-10 a 17	1.181	1.042
12	PN + UN	N + DN + M	-10 a 17	2.143	1.106
13	PN + PM + UM	N + DN + M	-10 a 17	1.286	1.012
14	PN + UM	N + DN + M	-10 a 17	1.591	1.053
15	UN + AN	N + DN + M	-10 a 17	1.919	1.080
16	AN	N + DN + M	-10 a 17	12.581	1.145

AN: ano de nascimento; CNS: Cartão Nacional de Saúde; DN: data de nascimento; M: mãe; MUN: município; N: nome; PM: código *Soundex* do primeiro nome da mãe; PN: código *Soundex* do primeiro nome; UM: código *Soundex* do último nome da mãe; UN: código *Soundex* do último nome.

Tabela 2

Combinação de pares do relacionamento dos bancos Sistema de Informação do Câncer de Mama (SISMAMA), Minas Gerais, Brasil, 2009 e 2010.

Passos	Número de pares	% (total = 1.194)
16, 9	1.151	96,40
16, 9, 12	1.185	99,24
16, 9, 12, 8	1.185	99,24
16, 9, 12, 1	1.188	99,50
16, 9, 12, 1, 15	1.188	99,50
16, 9, 12, 1, 2	1.188	99,50
16, 9, 12, 1, 5	1.188	99,50
16, 9, 1, 14	1.190	99,67
16, 9, 12, 1, 14, 11	1.192	99,83
16, 9, 1, 14, 11, 7	1.193	99,92
16, 9, 12, 1, 14, 11, 7, 13	1.194	100,00

Tabela 3

Ponto de corte dos passos e amostras do Relacionamento dos bancos do Sistema de Informação do Câncer de Mama (SISMAMA), Minas Gerais, Brasil, 2009 e 2010.

P	Ponto de corte			
	T	A1	A2	A3
1	9,84	10,96	10,64	10,69
2	9,76	3,00	9,93	9,87
3	-	-	-	-
4	6,54	1,00	7,06	6,54
5	6,54	2,00	7,06	6,54
6	6,54	1,00	6,75	6,69
7	6,56	2,00	6,75	6,44
8	6,54	1,00	6,24	6,54
9	6,54	3,00	6,65	6,70
10	6,54	1,00	7,06	6,54
11	6,69	1,00	6,72	6,24
12	6,56	2,00	6,73	6,70
13	6,56	1,00	6,29	6,69
14	6,56	2,00	6,29	6,44
15	6,54	3,00	6,70	6,56
16	6,54	14,00	7,06	7,06

P: passo; T: total de pares (n = 1.194); A1: amostra de 100 pares; A2: amostra de 200 pares; A3: amostra de 300 pares.

precision-recall (> 0,98). A maioria das amostras de 100 pares evidenciou altos valores para essas medidas, porém com escores mais baixos. Somente o passo cuja variável de blocagem foi ano de nascimento e de comparação nome, data de nascimento e nome da mãe (passo 16) apresentou baixa sensibilidade (*recall*), medida F e área sob a curva *precision-recall* (Tabelas 4 e 5).

Discussão

Este estudo demonstrou que é possível automatizar o relacionamento dos bancos de dados em saúde mantendo a acurácia do método, com base na inspeção manual de uma amostra retirada do próprio banco e posterior seleção de um escore de alto poder discriminatório na identificação de pares verdadeiros desta amostra.

Alguns trabalhos usaram escores para os limiares superior e inferior [4,5,6], indicando para a inspeção os pares com escores intermediários, porém isto pode não ser aplicável quando existe um grande números de pares a ser revisados. Desse modo, recomendamos amostragem aleatória com no mínimo 200 registros, inspeção manual, cálculo da sensibilidade (*recall*), do valor preditivo positivo (*precision*), da medida F, construção da curva *precision-recall* dos escores dessa amostra e posterior aplicação desse ponto de corte para a classificação automática dos pares do banco total.

Estudos [1,5,8] sobre acurácia do *linkage* encontraram medidas de qualidade semelhantes às do presente trabalho tanto para o total de cada passo quanto para as amostras de 200 e 300 pares: sensibilidade > 95%, valor preditivo positivo > 95% e acurácia > 97%.

Por meio da aplicação de três passos foi verificado que é possível encontrar quase a totalidade de pares verdadeiros e que a inclusão de mais passos agrega poucos pares. O estudo realizado por Girianelli et al. [4] propõe a utilização de 15 passos, contudo, dependendo do tamanho da base e da disponibilidade de tempo não é possível o uso de todas estas estratégias. As autoras referem ainda que com quatro passos é possível identificar a maioria dos pares verdadeiros e que ao utilizar a chave de blocagem primeiro nome, último nome, primeiro nome da mãe e último nome da mãe elas consegui-

Tabela 4

Recall e precision do ponto de corte dos passos e amostras do relacionamento dos bancos do Sistema de Informação do Câncer de Mama (SISMAMA), Minas Gerais Brasil, 2009 e 2010.

P	*Recall*				*Precision*			
	T	A1	A2	A3	T	A1	A2	A3
1	0,98	0,98	1,00	0,99	0,99	1,00	1,00	1,00
2	0,98	1,00	0,97	0,98	0,96	0,56	0,95	0,98
3	-	-	-	-	-	-	-	-
4	0,99	1,00	1,00	0,99	0,99	0,97	1,00	1,00
5	0,99	1,00	0,98	0,99	0,99	0,86	0,98	1,00
6	0,98	1,00	0,98	0,99	0,99	0,91	1,00	0,99
7	0,98	1,00	0,98	1,00	0,99	0,81	0,99	0,98
8	0,98	1,00	0,99	0,97	0,99	0,86	0,99	1,00
9	0,97	1,00	0,98	0,97	0,99	0,57	0,97	1,00
10	0,99	1,00	1,00	1,00	0,98	0,95	0,97	0,97
11	0,98	1,00	0,99	0,99	0,98	0,90	0,98	0,97
12	0,98	1,00	0,98	0,98	0,97	0,56	0,94	0,97
13	0,98	1,00	0,99	0,99	0,98	0,77	0,96	0,97
14	0,97	1,00	1,00	1,00	0,97	0,68	0,97	0,97
15	0,99	1,00	0,99	0,99	0,99	0,62	0,98	0,98
16	0,98	1,00	1,00	1,00	0,98	0,30	1,00	0,97

P: passo; T: total de pares (n = 1.194); A1: amostra de 100 pares; A2: amostra de 200 pares; A3: amostra de 300 pares.

Tabela 5

Acurácia, medida F e área sob a curva *precision-recall* (AUPRC) do ponto de corte dos passos e amostras do relacionamento dos bancos do Sistema de Informação do Câncer de Mama (SISMAMA), Minas Gerais, Brasil, 2009 e 2010.

P	Acurácia				Medida F				AUPRC			
	T	A1	A2	A3	T	A1	A2	A3	T	A1	A2	A3
1	0,99	0,99	1,00	0,99	0,99	0,99	1,00	0,99	0,99	0,99	1,00	0,99
2	0,99	0,89	0,98	0,99	0,97	0,71	0,96	0,98	0,99	0,42	0,98	0,98
3	-	-	-	-	-	-	-	-	-	-	-	-
4	0,99	0,97	1,00	0,99	0,99	0,98	1,00	0,99	0,99	0,92	1,00	0,99
5	0,99	0,87	0,97	0,97	0,99	0,92	0,98	0,99	0,99	0,79	0,99	0,99
6	0,97	0,91	0,99	0,99	0,99	0,95	0,99	0,99	0,99	0,81	0,99	0,99
7	0,98	0,82	0,98	0,98	0,99	0,89	0,98	0,99	0,99	0,70	0,99	0,99
8	0,97	0,86	0,99	0,97	0,99	0,92	0,99	0,98	0,99	0,79	0,99	0,99
9	0,97	0,73	0,98	0,99	0,98	0,72	0,98	0,98	0,99	0,46	0,99	0,99
10	0,97	0,95	0,98	0,97	0,99	0,97	0,98	0,98	0,99	0,87	0,99	0,99
11	0,97	0,90	0,98	0,97	0,98	0,94	0,99	0,98	0,99	0,81	0,99	0,99
12	0,97	0,58	0,96	0,98	0,97	0,72	0,96	0,98	0,99	0,41	0,99	0,99
13	0,97	0,77	0,97	0,97	0,98	0,87	0,98	0,98	0,99	0,62	0,99	0,99
14	0,97	0,70	0,98	0,98	0,97	0,81	0,98	0,98	0,99	0,63	0,99	0,99
15	0,99	0,72	0,98	0,99	0,99	0,77	0,98	0,99	0,99	0,62	0,99	0,99
16	0,99	0,95	1,00	0,99	0,98	0,46	1,00	0,98	0,99	0,32	1,00	0,99

P: passo; T: total de pares (n = 1.194); A1: amostra de 100 pares; A2: amostra de 200 pares; A3: amostra de 300 pares.

ram recuperar pares que não tinham sido identificados nos passos anteriores. No presente trabalho, o passo que usa as chaves de blocagem primeiro nome e ano de nascimento (passo 9) compartilha 97,2% de pares verdadeiros com o passo que usa as chaves primeiro nome, último nome, primeiro nome da mãe e último nome da mãe (passo 10).

O uso de maior número de variáveis na blocagem e pareamento como, por exemplo, nos passos 4, 6 e 10, aumenta o número de ligações verdadeiras (> 90%) e restringe as falsas (< 8%). Em contrapartida, estratégias poucos restritas, como nos passos 2 e 16, aumentam muito o número de pares formados e, consequentemente, eleva a quantidade de falsos pares.

A utilização de um identificador único como chave de blocagem contribui para que a maior parte ou a totalidade dos pares formados seja verdadeiro, como verificado no presente estudo que com o CNS (passo 3) formou 216 pares, todos verdadeiros. Os escores baixos apresentados por pares desse passo podem ser explicados pela incompletude de informações. Em um relacionamento entre banco de doadores de sangue e o Sistema de Informações sobre Mortalidade (SIM) foram encontrados dezesseis pares verdadeiros em baixos escores, sendo observado que eles apresentavam dados incompletos, principalmente para a variável nome [10]. Esse mesmo estudo obteve sensibilidade de 94%.

Variáveis importantes na identificação do indivíduo, tais como: nome, nome da mãe e data de nascimento, devem estar muito bem preenchidas nos bancos. Um estudo [5] sobre a definição de um ponto de corte para a identificação de pares verdadeiros em registros de câncer usou três variáveis na blocagem e duas no pareamento, encontrando nos relacionamentos realizados, especificamente para o câncer de mama, 47,5% e 78,8% pares falsos, aproximadamente. Destaca-se que a qualidade da informação e sua completude são importantes fatores que influenciam o processo de formação de pares e que o escore é influenciado pelas variáveis selecionadas para o pareamento [2]. Um elevado porcentual de registros sem informação de nome da mãe no Sistema de Informação de Nascidos Vivos (SINASC) resultou em um relacionamento com baixa sensibilidade [17].

Relacionamentos probabilísticos têm como objetivo obter o maior número de pares verdadeiros possíveis, mesmo que para isto sejam obtidos alguns falsos pares. Dessa forma, o interesse está principalmente na sensibilidade do método [9]. Na escolha do ponto de corte também pode ser considerada a máxima *precision* com a finalidade de minimizar a possibilidade de fazer correspondências falsas e, se necessário, aceitar algum nível de falhas para encontrar correspondências verdadeiras [10].

Atualmente, o grande volume de informações sociodemográficas, de saúde e de outras áreas, desperta a necessidade de agregação destes dados para que os gestores destes setores tenham conhecimento da realidade e possam definir prioridades. Assim, o relacionamento de bases de dados ganha destaque e importância como ferramenta de integração dos diversos sistemas de informação, tornando-se essencial a otimização destes processos.

As limitações do estudo são inerentes ao uso de bases de dados secundários devido à sua qualidade variável, podendo conter incompletudes e erros de digitação. Além disso, cabe assinalar que a utilização da revisão manual como padrão-ouro pode produzir erros em cenários em que os revisores não estejam preparados adequadamente ou estejam submetidos a rotinas fatigantes, o que pode ocasionar classificações equivocadas entre pares verdadeiros e falsos.

Destaca-se que o processo de limpeza das bases de dados e retirada de duplicidades e repetições, que antecede a etapa de relacionamento, é fundamental para que sejam mantidas no banco apenas as variáveis de interesse, um registro de cada indivíduo e para a eliminação de caracteres especiais, facilitando o processo de ligação de pares.

Os resultados apresentados dão suporte aos gestores de saúde e àqueles que utilizam o relacionamento de bancos de dados, para que com base no seu próprio banco identifiquem, valendo-se de uma amostra selecionada aleatoriamente, o escore com o melhor desempenho na identificação de pares verdadeiros e com o mínimo possível de perdas, o que permite automatizar o relacionamento dos bancos mantendo a acurácia do método.

Colaboradores

D. A. P. Duarte, M. C. Nogueira e M. T. Bustamante-Teixeira contribuíram na concepção e delineamento do estudo, análise e interpretação dos dados e redação do manuscrito; revisaram e aprovaram a versão final, e são responsável por todos os aspectos do trabalho. C. S. L. Corrêa e V. A. Fayer contribuíram na análise, interpretação dos dados e redação do manuscrito; revisaram e aprovaram a versão final, e são responsável por todos os aspectos do trabalho.

Informações adicionais

ORCID: Daniela de Almeida Pereira Duarte (0000-0002-0177-3676); Camila Soares Lima Corrêa (0000-0003-4315-7123); Vívian Assis Fayer (0000-0003-2529-5039); Mário Círio Nogueira (0000-0001-9688-4557); Maria Teresa Bustamante-Teixeira (0000-0003-0727-4170).

Agradecimentos

Agradecemos aos revisores pelas preciosas sugestões, que contribuíram para o aprimoramento do artigo. À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG): CDS-APQ 03809-17 pelo financiamento.

Referências

1. Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida. Cad Saúde Pública 2007; 22:2249-52.
2. Silveira DP, Artmann E. Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática. Rev Saúde Pública 2009; 43:875-82.
3. Caetano MC. Acurácia do relacionamento probabilístico na avaliação da alta complexidade em cardiologia. Rev Saúde Pública 2011; 45:269-75.
4. Tomazelli JG, Girianelli VR, Silva GA. Estratégias usadas no relacionamento entre Sistemas de Informações em Saúde para seguimento das mulheres com mamografias suspeitas no Sistema Único de Saúde. Rev Bras Epidemiol 2018; 21:e180015.
5. Peres SV, Tanaka LF, Latorre MRDO, Almeida MF, Coeli CM, Michels FAS. Determinação de um ponto de corte para a identificação de pares verdadeiros pelo método probabilístico de linkage de base de dados. Cad Saúde Colet (Rio J.) 2015; 22:428-36.
6. Romero ROG, Sá LD, Ribeiro CMC, Villa TCS, Nogueira JDA. Subnotificação de casos de tuberculose a partir da vigilância do óbito. Rev Eletrônica Enferm 2016; 18:e1161.
7. Camargo Jr. KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cad Saúde Pública 2005; 16:439-47.
8. Fonseca MGP, Coeli CM, Lucena FFA, Veloso VG, Carvalho MS. Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database. Cad Saúde Pública 2010; 26:1431-8.
9. Harron KL, Doidge JC, Knight HE, Gilbert RE, Goldstein H, Cromwell DA, et al. A guide to evaluating linkage quality for the analysis of linked data. Int J Epidemiol 2017; 46:1699-710.
10. Capuani L, Bierrenbach AL, Abreu F, Takecian PL, Sabino EC. Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. Cad Saúde Pública 2014; 30:1623-32.
11. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015; 10:e0118432.
12. Boyd JH, Guiver T, Randall SM, Ferrante A M, Semmens J B, Anderson P, et al. A simple sampling method for estimating the accuracy of large scale record linkage projects. Methods Inf Med 2016; 55:276-83.
13. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. J Biomed Inform 2012; 45:165-72.
14. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. https://link.springer.com/content/pdf/10.1007%2F978-3-642-40994-3_29.pdf (acessado em 02/Jun/2019).
15. Camargo Junior K, 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.
16. Girianelli VR, Thuler LCS, Silva GA. Qualidade do sistema de informação do câncer do colo do útero no Estado do Rio de Janeiro. Rev Saúde Pública 2009; 43:580-8.
17. da Matta Coutinho RG, Coeli CM, Faerstein E, Chor D. Sensitivity of probabilistic record linkage for reported birth identification: Pró-Saúde Study. Rev Saúde Pública 2008; 42:1097-100.

Abstract

The objective was to test and assess the accuracy of a scoring method in probabilistic data linkage in order to enable automatic identification of true matches, dispensing with the manual inspection stage. Accuracy study using data from the Breast Cancer Information System (SISMAMA) base in Minas Gerais State, Brazil, from 2009 and 2010. After cleaning and standardization, a 16-step probabilistic linkage of the 2009 and 2010 databases was performed, where each step was inspected manually to obtain a gold standard. Samples were then selected, inspected, and assessed to calculate the method's accuracy in selecting true matches. All the steps and samples with 200 and 300 matches showed high sensitivity (recall) > 0.97, high positive predictive value (precision) > 0.95, high accuracy (> 0.97) and F measure (> 0.96), and high area under the curve precision-recall (> 0.98). The sample with 100 matches showed high values for these measures, but with low scores. Of the 16 steps assessed, the combined use of only three was sufficient to identify 99.24% of the true matches in the total database. The proposed method allows automatically linking databases, maintaining the method's accuracy. It facilitates the use of probabilistic linkage in health services, especially for health surveillance and management.

Health Information System; Systems Integration; Data Accuracy

Resumen

El objectivo fue robar y evaluar la exactitud de un método para la selección de una puntuación, en la relación probabilística de bancos de datos, de forma que sea viable la automatización de la identificación de pares verdaderos, eximiendo la etapa de revisión manual. Estudio de precisión, utilizando datos del Sistema de Información del Cáncer de Mama (SISMAMA) de Minas Gerais, Brasil, de 2009 y 2010. Tras el proceso de limpieza y estandarización, se realizó la relación probabilística de los bancos 2009 y 2010, utilizando 16 pasos, donde cada paso se revisó manualmente para obtener un patrón-oro. Posteriormente, se seleccionaron muestras que fueron revisadas y evaluadas para calcular la precisión del método de selección de los pares verdaderos. Todos los pasos y muestras con 200 y 300 pares presentaron una alta sensibilidad (recall) *> 0,97, un alto valor predictivo positivo* (precision) *> 0,95 y exactitud alta (> 0,97), medida F (> 0,96) y el área bajo la curva* precision-recall *(> 0,98). La muestra con 100 pares evidenció altos valores para estas medidas, aunque con puntuaciones más bajas. De los 16 pasos evaluados, el uso de solo tres de forma combinada fueron suficientes para identificar 99,24% de los pares verdaderos en el banco total. El método propuesto permite automatizar la relación de las bases de datos, manteniendo la precisión del método. Facilita la utilización de la relación probabilística en el ámbito de los servicios de salud, especialmente para vigilancia y gestión en salud.*

Sistemas de Información en Salud; Integración de Sistemas; Exactitud de los Datos

Recebido em 05/Abr/2019
Versão final reapresentada em 21/Ago/2019
Aprovado em 23/Ago/2019