RSP Revista de Saúde Pública

http://www.rsp.fsp.usp.br/

Acurácia das técnicas de relacionamento probabilístico e determinístico: o caso da tuberculose

Gisele Pinto de Oliveira[I], **Ana Luiza de Souza Bierrenbach**[II], **Kenneth Rochel de Camargo Júnior**[III], **Cláudia Medina Coeli**[IV], **Rejane Sobrino Pinheiro**[IV]

[I] Programa de Pós-Graduação em Saúde Coletiva. Instituto de Estudos em Saúde Coletiva. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil
[II] Instituto de Ensino e Pesquisa. Hospital Sírio-Libanês. São Paulo, SP, Brasil
[III] Instituto de Medicina Social. Universidade do Estado do Rio de Janeiro. Rio de Janeiro, RJ, Brasil
[IV] Instituto de Estudos em Saúde Coletiva. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil

RESUMO

OBJETIVO: Analisar a acurácia das técnicas determinística e probabilística para identificação de registros duplicados de tuberculose, assim como as características dos pares discordantes.

MÉTODOS: Foram analisados todos os registros de tuberculose no período de 2009 a 2011 do estado do Rio de Janeiro. Foi desenvolvido algoritmo para relacionamento determinístico, usando conjunto de 70 regras, a partir da combinação de fragmentos das variáveis-chave com ou sem modificações (*Soundex* ou *substring*). Cada regra era formada por três ou mais fragmentos. Para a abordagem probabilística, foi necessário estabelecer ponto de corte para o escore, acima do qual os *links* seriam classificados automaticamente como pertencentes ao mesmo indivíduo. O ponto de corte foi obtido por meio do relacionamento da base de dados Sistema de Informação de Agravos de Notificação – Tuberculose com ela mesma, posterior revisão manual e curvas ROC e *precision-recall*. Foram calculadas a sensibilidade e especificidade para análise de acurácia.

RESULTADOS: A acurácia variou de 87,2% a 95,2% para sensibilidade e 99,8% a 99,9% para especificidade para as técnicas probabilística e determinística, respectivamente. A presença de valores faltantes para as variáveis-chave e o baixo percentual da medida de similaridade para o nome e data de nascimento foram os principais responsáveis pela não identificação dos registros do mesmo indivíduo pelas técnicas utilizadas.

CONCLUSÕES: As duas técnicas apresentam alta concordância para a classificação como par. Apesar de a técnica determinística ter identificado mais registros duplicados que a probabilística, a segunda recuperou registros não identificados pela primeira. A necessidade e a experiência do usuário devem ser consideradas para a escolha da técnica a ser utilizada.

DESCRITORES: Tuberculose, epidemiologia. Confiabilidade dos Dados. Sensibilidade e Especificidade. Vigilância Epidemiológica, estatística & dados numéricos.

Correspondência:
Gisele Pinto de Oliveira
Instituto de Estudos em Saúde Coletiva – IESC
Praça Jorge Moreira Machado, 1000 Cidade Universitária
21941-598 Rio de Janeiro, RJ, Brasil
E-mail: giselepoliveira@gmail.com

Recebido: 15 abr 2015
Aprovado: 1 set 2015

Como citar: Oliveira GP, Bierrenbach ALS, Camargo Jr KR, Coeli CM, Pinheiro RS. Acurácia das técnicas de relacionamento probabilístico e determinístico: o caso da tuberculose. Rev Saude Publica. 2016;50:49.

INTRODUÇÃO

O Brasil, assim como outros países, possui grande volume de dados do Setor Saúde coletado por meio de sistemas de informações nacionais e disponíveis em bases de dados distintas. O relacionamento entre bases de dados visa identificar se dois ou mais registros dizem respeito à mesma entidade, em geral, mesmo indivíduo. Essa técnica é usada para identificar duplicidades em um mesmo arquivo ou entre dois ou mais arquivos que precisam ter suas informações consolidadas em banco de dados único[12]. A ausência do campo identificador unívoco nas bases de dados impede a identificação direta de registros de um mesmo indivíduo em bases de dados distintas, necessitando da aplicação de algoritmos de encadeamento mais sofisticados que se baseiam na combinação de variáveis de identificação. Os relacionamentos probabilístico e determinístico são técnicas empregadas para esse fim[6,10,18,19].

O relacionamento probabilístico de registros usa funções de comparação aproximadas. Pesos diferentes são atribuídos a cada campo com base no seu poder de discriminação e vulnerabilidade ao erro. O relacionamento determinístico utiliza funções de comparação exatas e classificação baseada em regras desenvolvidas a partir do conhecimento de especialistas[6]. Rotinas computacionais específicas precisam ser desenvolvidas para cada problema. A baixa qualidade dos dados, como presença de dados faltantes e erros de digitação, pode contribuir para erros de pareamento das variáveis, o que torna importante avaliar a acurácia das técnicas de relacionamento de bases de dados[16].

O Ministério da Saúde é responsável por desenvolver, gerenciar e armazenar dados dos sistemas de informação nacionais em saúde. Embora os sistemas não se interliguem automaticamente, é possível relacionar essas bases de dados, pois existem variáveis de identificação nominal com elevado poder discriminatório, especialmente quando usadas de forma combinada, que são padronizadas nos sistemas. Os casos de tuberculose (TB) são registrados compulsoriamente em um sistema de informação nacional para notificação de doenças (Sistema de Informação de Agravos de Notificação – Sinan-TB)[a]. O sistema de vigilância da TB permite que cada indivíduo tenha diferentes entradas já que, pelas normas vigentes no País, é necessário notificar os casos de recidiva e reingresso após abandono. Entretanto, existem registros duplicados indevidamente que dizem respeito ao mesmo episódio da doença e que necessitam ser reconhecidos, para serem corretamente eliminados[1-3,b].

Poucos estudos comparam a acurácia de processos de relacionamento entre bases de dados secundárias[15]. Além disso, parece não existir na literatura científica orientação de como escolher entre técnicas existentes e comparação de seus resultados.

O objetivo desse trabalho foi analisar a acurácia das técnicas determinística e probabilística para a identificação de registros duplicados de tuberculose, assim como as características dos pares discordantes.

MÉTODOS

Foram utilizadas as bases de dados do Sinan-TB, do período de 2009 a 2011, do estado do Rio de Janeiro, disponibilizada pela Secretaria de Estado de Saúde. As notificações de TB cujo encerramento deveu-se à mudança de diagnóstico foram excluídas.

Foi realizada uma etapa comum de pré-processamento dos dados, visando corrigir erros e padronizar o teor das variáveis-chave (nome; nome da mãe; data de nascimento; logradouro e bairro) utilizadas em cada técnica. As transformações de dados incluíram: remoção de sinais de pontuação, acentos, espaços em branco repetidos e preposições; conversão das letras para maiúsculo; remoção de números das variáveis que devem ser exclusivamente compostas de letras e vice-versa, remoção de termos que indicam a falta de informação (não sei, desconhecido, entre outros), substituição de letras duplas por uma única, padronização de formatos de data, padronização de termos usados em logradouros ("R." foi substituído por "Rua", "Av." por "Avenida" etc.).

[a] Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica. Sistema de Informação de Agravos de Notificação (SINAN): normas e rotinas. 2.ed. Brasília (DF); 2007.

[b] Ministério da Saúde, Secretaria de Vigilância em Saúde. Manual de recomendações para o controle da tuberculose no Brasil. Brasília (DF); 2011. (Série A. Normas e Manuais Técnicos).

Posteriormente, foram geradas variáveis novas a partir do nome da mãe e endereço padronizadas, realizando processos de: 1) *Parsing* (separação dos fragmentos em primeiro nome, segundo nome, e assim por diante) e 2) *Substringing* (partes do fragmento como "Maria" → "Mari"; "Oliveira" → "Veira"). Cada variável nova foi também transformada em outra contendo o seu código *Soundex*[8]. Esse recurso visa transformar o texto em um código fonético, visando homogeneizar pequenas diferenças de escrita (resultantes de erros de pronunciação ou de entendimento por parte de quem o escreveu), uso ou não de consoantes duplas, entre outros.

A abordagem probabilística envolve processos de padronização, blocagem e formação dos *links* (par de registros a serem comparados), aplicação de algoritmos de comparação e geração de escore de similaridade, definição de limiares para a classificação dos *links* em pares verdadeiros, não pares e pares duvidosos, revisão manual dos pares duvidosos e remoção de duplicidades[7,c].

O ponto de corte do escore, acima do qual os *links* foram classificados como pertencentes ao mesmo indivíduo[5,c], foi obtido por meio do relacionamento da base de dados Sinan-TB com ela mesma, por um período menor: 2010-2011. Utilizou-se apenas uma estratégia de blocagem, mais sensível, com o *Soundex* do primeiro nome e sexo, para possibilitar o maior número de combinações de registros. O arquivo de *links* foi revisado manualmente com o objetivo de classificá-los como pares ou não pares, ou seja, como pertencentes ou não ao mesmo indivíduo, respectivamente. A distribuição dos *links* classificados como não pares variou de -7,5 a 19,7, enquanto para os pares variou de 6,9 a 34,9. Foi estabelecido o ponto de corte de 18,3 a partir da análise exploratória das distribuições dos escores e pela inspeção das curvas ROC e *precision-recall*. Foi utilizada a biblioteca ROCR do pacote estatístico R versão 2.15.3[17].

Para a identificação de registros duplicados, a mesma estratégia de blocagem foi utilizada. As variáveis do relacionamento foram nome, nome da mãe e data de nascimento. As duas primeiras foram comparadas utilizando a distância de Levenshtein e a terceira, usando um algoritmo exato[6]. Para cada *link*, um escore dado pelo peso composto foi calculado pela soma do acordo ou do desacordo para cada variável a ser comparada[9,c]. Foi aplicado o escore de 18,3 para classificação automática e formação dos grupos de registros prováveis de pertencerem ao mesmo indivíduo. Foi utilizado o *software OpenReclink*, amplamente empregado na área de saúde, o qual possui uma rotina específica para identificação de duplicidades[c]. O resultado dessa rotina é arquivo de entrada acrescido de um escore e do código identificador do grupo. Dessa forma, não são conhecidos quais *links* foram formados. Por exemplo, para um grupo formado pelos registros A, B, C e D, não se sabe se todos os registros combinaram entre si no processo de relacionamento ou se A combinou com B e C, e B combinou com D. Foi elaborada uma estratégia para reconstituição desses *links*: quando não havia igualdade de escore entre todos os registros de um mesmo grupo, optou-se por considerar o registro com escore mais alto do grupo para formar o *links* com o registro de escore diferente.

Foi desenvolvido um algoritmo para relacionamento determinístico no Stata 12.0, baseado em 70 regras, para identificar os grupos de registros do mesmo indivíduo e classificá-los como par ou não par. Tais regras foram formadas a partir das variáveis-chave, bem como daquelas criadas durante o pré-processamento. Ainda, foram utilizados, em algumas regras, os códigos *Soundex* de forma concatenada (por exemplo, "Maria Antonia Santos" foi representado por M600A535S532). Em cada regra, o conteúdo de três ou mais variáveis foi comparado de forma exata, na presença ou não de uma série de restrições de dados (Tabela 1).

A variável nome foi utilizada em quase todas as regras, dado seu elevado poder de discriminação. A data de nascimento foi utilizada, em algumas regras, com a informação de dia, mês e ano, e em outras, um ou mais dos seus componentes foram utilizados de forma combinada. As variáveis código do município de residência ou de notificação, sexo, data de notificação, data de desfecho, número de notificação, código da unidade de notificação, assim como outras relativas ao local de residência (bairro, telefone, CEP e número do prédio) foram utilizadas na forma original. Bairro foi a única dessas variáveis que recebeu pré-processamento.

[c] Camargo Jr KR, Coeli CM. OpenReclink: guia do usuário [citado 2015 jun 5]. Disponível em: http://reclink.sourceforge.net/

Tabela 1. Exemplos de algumas regras utilizadas no algoritmo sequencial.

Regra	Nome do paciente	Nome da mãe	Data de nascimento	Endereço	Restrições utilizadas
1	Exato	Exato	Exato	-	Sem valores faltantes Nome completo do paciente com pelo menos 15 caracteres de comprimento
2	Mesmo *Soundex* (do nome completo)	Exato	Exato	-	Sem valores faltantes *Soundex* composto de 3 ou + pedaços Sem recém-nascidos ou gêmeos
3	Mesmo *Soundex* (do nome completo)	Mesmo *Soundex* (do nome completo)	Mesmos dia e ano	-	Sem valores faltantes Somente considerando nomes incomuns
4	Mesmos quatro caracteres iniciais para primeiro e segundo nomes + *Soundex* exato para sobrenome	Mesmo *Soundex* (primeiro nome e sobrenome)	Exato	Exato	Sem valores faltantes Sem recém-nascidos ou gêmeos

As regras foram dispostas de forma sequencial: das que usavam menos variáveis, com maior poder de discriminação, para as que usavam combinação de muitas variáveis, mas que cada uma isoladamente tivesse menor poder de discriminação. Entende-se por poder de discriminação a capacidade de uma variável conseguir de forma isolada discriminar um indivíduo.

A inclusão das regras buscou aumentar a sensibilidade do algoritmo, sem perder especificidade. A cada nova regra, novos grupos eram encontrados ou novos registros adicionados aos grupos existentes, mesmo que esses não fizessem par com todos os registros já identificados do grupo. As regras que geraram grupos incorretos foram modificadas ou finalmente descartadas após extensa revisão manual.

Várias restrições de dados foram impostas em muitas regras. A restrição universal para todas as regras era a de que os grupos somente poderiam ser construídos com registros que não possuíssem variáveis com valores faltantes. Além disso, para certas regras que usaram uma combinação de variáveis com menor poder de discriminação, o conteúdo de algumas dessas variáveis tinha que ter um comprimento mínimo de caracteres. Do mesmo modo, em algumas regras, o código *Soundex* tinha que ter um número mínimo de blocos (Maria Antonia Santos → M600-A535-S532 → 3 blocos).

A raridade dos nomes também foi considerada em algumas regras. Para isso, três bases de dados separadas foram criadas indicando a frequência de aparecimento do primeiro, segundo e último nome no banco de dados do Sistema de Informações sobre Mortalidade (SIM) do Brasil dos anos de 2008 a 2010. Nomes presentes no banco de dados de TB mas ausentes no SIM foram considerados raros. Registros contendo os mesmos primeiro, segundo ou últimos nomes considerados raros foram reunidos em grupos, mesmo que houvesse informação incompleta em algumas das outras variáveis usadas na regra. No entanto, não foram considerados como pertencentes ao mesmo grupo registros que tinham nomes raros iguais, porém com datas de nascimento distantes mais de 20 anos, pois poderiam se referir a pai e filho, por exemplo. Essa possibilidade tinha de ser considerada, especialmente, por se tratar de uma doença infecciosa cuja principal transmissão é intradomiciliar. A definição de

raridade variou de 500 a 1.000 para a frequência do nome, conforme o poder discriminatório das outras variáveis da regra.

Foi realizada revisão manual dos *links* classificados como pares pelas duas técnicas e daqueles que discordavam quanto à classificação. A classificação da revisão manual foi considerada o padrão-ouro para os cálculos de sensibilidade e especificidade.

Foram analisados os seguintes indicadores para traçar o perfil dos registros com classificação discordante entre as técnicas:

1. Percentual de valores faltantes para as variáveis nome, nome da mãe, sexo, data de nascimento e logradouro;
2. Mediana do escore dos registros duplicados identificados pela técnica probabilística e não identificados pela determinística;
3. Percentual de registros do mesmo grupo identificados pela técnica probabilística que apresentaram diferença na variável sexo;
4. Mediana da medida de similaridade para as variáveis nome, nome da mãe e data de nascimento dos registros discordantes (calculada somente para registros sem valores faltantes).

A distância de Levenshtein (DL) deve ser analisada considerando o tamanho da cadeia de caracteres, já que cadeias longas têm mais chance de gerar valores maiores[19]. Uma medida de similaridade (MS) foi obtida subtraindo-se a DL do comprimento da cadeia completa mais longa e dividindo-se também por este valor. O resultado foi multiplicado por 100 para obter-se a porcentagem e facilitar a análise do resultado.

Para os grupos com três ou mais registros, o *link* de registros usado no cálculo da MS foi formado pelo registro que apresentou classificação discordante entre as técnicas e o registro que, no processo probabilístico, apresentou o maior escore do grupo. Essa alternativa foi adotada supondo-se que o registro associado ao maior escore tenderia a ser o grafado mais corretamente. Os registros que apresentaram diferença no sexo foram excluídos da análise da MS, uma vez que não tiveram oportunidade de serem relacionados pela abordagem probabilística, porque pertenciam a blocos lógicos distintos. Foi utilizado o *software* Stata 12.0 para as análises dos registros.

O projeto foi aprovado pelo Comitê de Ética em Pesquisa do Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro (Processo 114.604 de 3/10/2012).

RESULTADOS

A base de dados do Sinan-TB continha 43.825 registros. A Tabela 2 apresenta o percentual de registros simples ou que formam grupos de dois até 10 registros identificados pelas técnicas determinística e probabilística. Na determinística, 78,7% eram registros simples e, na probabilística, 80,5%.

Na técnica determinística, 21,3% dos registros foram classificados como pares e, na probabilística, 19,5%. Apresentaram classificação discordante 1.812. Um total de 527 registros foi classificado como par pela técnica probabilística e pela determinística não, e 1.285 registros formaram par pela determinística e pela probabilística não. A revisão manual posterior, realizada para a formação do padrão-ouro, não evidenciou alteração de classificação para o grupo de registros classificados como pares pelas duas técnicas (Tabela 3).

Na análise da acurácia, os valores de sensibilidade e especificidade para a técnica determinística foram de 95,3% e 99,9%, respectivamente. Para a técnica probabilística, 87,2% e 99,8%, respectivamente (Tabela 4).

Tabela 2. Número e percentual de registros simples ou por grupos segundo as técnicas determinística e probabilística. Sinan-TB, 2009-2011.

Registros por grupo (n)	Técnica determinística		Técnica probabilística	
	Casos (n)	%	Casos (n)	%
1	34.506	78,7	35.266	80,5
2	6.944	15,8	6.546	14,9
3	1.502	3,4	1.280	2,9
4	548	1,3	468	1,1
5	215	0,5	160	0,4
6	66	0,2	60	0,1
7	35	0,1	35	0,1
9	9	0	-	-
10	-	-	10	0
Total	43.825	100	43.825	100

Tabela 3. Análise de concordância das técnicas utilizadas.

Técnica probabilística	Técnica determinística				Total
	Não par		Par		
	n	%	n	%	
Não par	33.980	96,4	1.285	3,6	35.265
Par	527	6,2	8.033	93,8	8.560
Total	34.507	78,7	9.318	21,3	43.825

Tabela 4. Análise de sensibilidade e especificidade das técnicas utilizadas.

Padrão	Total	Técnica determinística		Técnica probabilística	
		Par	Não par	Par	Não par
Par	9.741	9.283	458	8.491	1.250
Não par	34.084	35	34.049	69	34.015
Total	43.825	9.318	34.507	8.560	35.265
Sensibilidade (IC95%)		95,3	(94,8–95,7)	87,2	(86,5–87,8)
Especificidade (IC95%)		99,9	(99,8–99,9)	99,8	(99,7–99,8)

Dos registros que a técnica determinística classificou como par e a probabilística não, nenhum apresentou valores faltantes nas variáveis nome e sexo, 5,3% possuíam valores faltantes para nome da mãe, 6,3% na data de nascimento e 0,5%, no logradouro. Dez por cento dos registros apresentaram valores faltantes em pelo menos uma dessas variáveis analisadas. A mediana do escore dos registros classificados como par pela técnica probabilística e como não par pela determinística foi de 24,2, variando de 20,3 a 32,3. Nenhum registro com informação ignorada foi encontrado para as variáveis nome e sexo. Para o nome da mãe, 14,4% dos registros possuíam valores faltantes; para a data de nascimento, 11,0%; e para o logradouro, 3,8%. Mais de um quarto dos registros apresentaram valores faltantes em pelo menos uma das variáveis analisadas.

Dos 733 *links* classificados como par pela técnica determinística e não par pela probabilística, 15,7% apresentaram preenchimento distinto na variável sexo. Aproximadamente 26,0% apresentaram MS menor que 70,0% para a variável nome, 26,6%, para o nome da mãe e 8,3%, para data de nascimento. Para os *links* classificados como par apenas pelo probabilístico, 16,6% apresentou MS menor que 70,0% para a variável nome da mãe e 10,6%, para a data de nascimento. Quanto à mediana da MS, a maior diferença entre os grupos foi para a variável nome (81,6 para os *links* pares pelo determinístico e 100 para os *links* pares pelo probabilístico) (Tabela 5).

Tabela 5. Características dos registros com classificação discordante no uso das técnicas determinística e probabilística.

Características dos registros	Determinístico classificou como par e probabilístico não (N = 1.285)		Probabilístico classificou como par e determinístico não (N = 527)	
			Mediana	IC95%
Escore	-	-	24,2	20,3–32,3
	n	%	n	%
Valor faltante para sexo	0	0	0	0
Valor faltante para nome	0	0	0	0
Valor faltante para nome da mãe	68	5,3	76	14,4
Valor faltante para data de nascimento	81	6,3	58	11,0
Valor faltante para endereço	7	0,5	20	3,8
Combinado: nome da mãe ignorado ou data de nascimento ignorada ou endereço ignorado	129	10,0	141	26,7
Características dos *links*	Determinístico classificou como par e probabilístico não[a,b] (N = 733)		Probabilístico classificou como par e o determinístico não[a,b] (N = 293)	
	n	%	n	%
Diferença no sexo	115	15,7	0	0
Medida de similaridade para o nome menor que 70,0%	160	25,9	0	0
Medida de similaridade para o nome da mãe menor que 70,0%	147	26,6	36	16,6
Medida de similaridade para data de nascimento menor que 70,0%	45	8,3	25	10,6
	Mediana	IC95%	Mediana	IC95%
Medida de similaridade para o nome x 100	81,6	69,4–94,4	100	95,0–100
Medida de similaridade para o nome da mãe x 100	91,3	68,2–100	94,4	85,7–100
Medida de similaridade para a data de nascimento	100	10,0–100	87,5	75,0–100

[a] Para calcular a distância de Levenshtein e avaliar a diferença de sexo entre os registros, foi necessário comparar os registros do grupo de registros do mesmo paciente. Comparar os registros discordantes: (i) quando os dois discordantes foram identificados por apenas uma das técnicas, o cálculo era feito entre eles; (ii) quando eram registros identificados pelas duas técnicas e apenas um deles não era identificado por uma das técnicas, o cálculo foi feito comparando o registro não identificado com o registro de maior escore do grupo.
[b] Para esse cálculo, o grupo de registros que foram blocados por sexo ou que tinha preenchimento em branco em uma das variáveis foram excluídos.

DISCUSSÃO

As duas técnicas apresentaram alta concordância na classificação dos registros como par. Entretanto, o uso da técnica determinística recuperou 8,8% mais registros que a probabilística, ainda que esta última tenha identificado registros não identificados pela primeira. Embora a especificidade das duas técnicas tenha sido semelhante, a sensibilidade foi maior para a determinística, corroborando os achados de outros estudos utilizando base de dados do Sinan[8,15]. Valores das medidas de sensibilidade e especificidade encontrados para a técnica probabilística neste estudo estão de acordo com os apresentados em outros estudos[9,12,13,16].

A presença de variáveis com valores faltantes e MS menor ou igual a 70,0% para variáveis-chave foi o principal motivo da não identificação dos pares pelas duas técnicas. Adicionalmente, a blocagem por sexo para a técnica probabilística foi também responsável pela não formação

de alguns grupos. Retirar a variável sexo da blocagem inicial para o estabelecimento do ponto de corte poderia minimizar esse problema, mas aumentaria muito o volume de *links* a serem revisados manualmente, pois a chave de blocagem se tornaria ainda mais sensível. Por outro lado, a revisão manual seria feita apenas uma vez e o ponto de corte estabelecido seria aplicado em trabalhos com a mesma base de dados. Embora não haja valores faltantes na variável nome, ainda é significante a falta de preenchimento das variáveis nome da mãe e data de nascimento no Sinan-TB[2]. A melhoria da qualidade dos dados da TB é um constante desafio no Brasil[1,13,17]. Esse problema é refletido nos indicadores epidemiológicos e de desempenho, mascarando a real situação da TB no País, o que enviesa as análises necessárias para a tomada de decisão e elaboração de novas estratégias de controle[1,2,11,14].

Variáveis que compunham o endereço foram usadas como critério auxiliar na classificação dos registros na abordagem determinística. Mesmo que essa variável seja utilizada com cautela, já que os pacientes podem mudar de endereço entre uma notificação e outra, ela auxilia na decisão. No relacionamento probabilístico entre bases de dados distintas, essas variáveis podem ser utilizadas durante a revisão manual para apoiar a classificação. Neste estudo, optamos por não fazer a revisão manual após a aplicação da rotina para identificação de duplicidades. A abordagem determinística considerou no algoritmo tratamento diferenciado para nomes comuns. Não foi analisada a influência dessas estratégias na recuperação de pares, mas acredita-se que tenham influenciado a classificação. A estratégia de pareamento por frequência (informação sobre raridade do nome) poderia ser incorporada ao *software* de relacionamento probabilístico, minimizando a ocorrência de falso positivo em escores elevados em função da presença de homônimos.

A mediana de escore elevada para os registros classificados como pares pela técnica probabilística e não pares pela determinística indica que variáveis-chave semelhantes entre si não foram identificadas. Dessa forma, fica demonstrada a necessidade de aprimorar a rotina determinística.

O algoritmo determinístico utilizou *software* estatístico comercial, cuja interface está na língua inglesa e requer conhecimento prévio do usuário quanto à linguagem de programação utilizada pelo *software*. Além disso, a adaptação do algoritmo para bases diferentes das do Sinan-TB está condicionada ao conhecimento dessas bases e à experiência em programação. Transcrevê-lo usando *software* livre pode facilitar a disseminação. A técnica probabilística utilizou *software* nacional livre, que também requer conhecimento prévio para seu uso. Entretanto, como a rotina de identificação de duplicidades do *OpenReclink* permite a classificação automática dos grupos, seu uso é facilitado[4,c].

O estudo não mensurou o tempo e as dificuldades na utilização de cada técnica. No entanto, tanto a confecção do algoritmo determinístico quanto a revisão manual para o estabelecimento do ponto de corte no probabilístico demandaram grande tempo e experiência dos pesquisadores.

O Sinan dispõe de rotinas intrínsecas específicas para identificação e tratamento de registros duplicados[15]. Entretanto, seu algoritmo utiliza o nome sem que essa variável tenha passado por correção prévia, reduzindo sua eficiência. Além disso, apresenta como limitação o grande volume de registros, ausência de retroalimentação da esfera nacional às demais esferas e necessidade de boa qualidade e velocidade da rede de computadores locais. O elevado número de notificações anuais requer que o tratamento das duplicidades na esfera nacional seja realizado "fora" do Sinan, necessitando da aplicação de técnicas de relacionamento a cada transferência dos dados entre os níveis informatizados[a].

A definição de qual técnica utilizar para identificar registros duplicados no Sinan-TB depende do objetivo e da necessidade dos usuários. Como a busca pela perfeição do método é uma premissa no âmbito da pesquisa, a técnica determinística pode ser utilizada inicialmente por recuperar a maior parte de *links* e, em seguida, a probabilística, para recuperar os *links* que não foram encontrados na determinística. Na rotina dos programas de controle da tuberculose nas três esferas de governo, a remoção de duplicidades é executada diversas vezes por ano,

demandando grande tempo dos profissionais envolvidos. O uso da técnica determinística seria o indicado, desde que os pré-requisitos para sua utilização estejam satisfeitos. Caso contrário, indica-se a probabilística, e, sendo necessário torná-la mais sensível, pode-se realizar revisão manual dos grupos que tiveram escore abaixo do ponto de corte estabelecido, com o intuito de recuperar novos *links* não identificados anteriormente. Já para a construção de novas rotinas determinísticas e aprimoramento das existentes, a técnica probabilística, por ser genérica, deveria ser usada como estratégia inicial, visando ao acúmulo de conhecimento a ser incorporado no desenvolvimento da determinística. Esse acúmulo de conhecimento é importante, dado que a técnica determinística pouco específica pode gerar resultados bem menos acurados que os obtidos pela probabilística. Para Grannis et al., embora a sensibilidade de uma técnica determinística possa atingir 100%, ela pode diminuir consideravelmente para dados com diferentes características de identificação (nomes étnicos, por exemplo), considerando, portanto, a probabilística como mais eficiente[9]. A técnica probabilística, por outro lado, pode ainda aumentar sua acurácia, se incorporar novas variáveis ou estratégias para melhoria do processo de comparação e classificação dos *links*.

O Sinan sofre atualizações rotineiramente para atender às necessidades dos usuários e do sistema de vigilância nacional[a]. Este trabalho contribui para que novas estratégias de decisão possam ser discutidas e incorporadas na rotina de remoção de duplicidades. Além disso, possibilita subsidiar pesquisadores e profissionais do Sistema Único de Saúde na escolha de qual técnica de relacionamento utilizar. Face às grandes contribuições e bons resultados gerados pelo uso do relacionamento de bases de dados no setor saúde, é importante que outros estudos de acurácia de técnicas de relacionamento sejam realizados no Brasil.

REFERÊNCIAS

1. Bartholomay P, Oliveira GP, Pinheiro RS, Vasconcelos AMN. Melhoria da qualidade das informações sobre tuberculose a partir do relacionamento entre bases de dados. *Cad Saude Publica*. 2014;30(11):2459-70. DOI:10.1590/0102-311X00116313
2. Bierrenbach AL, Stevens AP, Gomes ABF, Noronha EF, Glatt R, Carvalho CN, et al. Efeito da remoção de notificações repetidas sobre a incidência da tuberculose no Brasil. *Rev Saude Publica*. 2007;41 Supl 1:67-76. DOI:10.1590/S0034-89102007000800010
3. Bierrenbach AL, Oliveira GP, Codenotti S, Gomes AB, Stevens AP. Duplicates and misclassification of tuberculosis notification records in Brazil, 2001–2007. *Int J Tuberc Lung Dis*. 2010;14(5):593-99.
4. Camargo Jr KR, Coeli CM. Going open source: some lessons learned from the development of OpenRecLink. *Cad Saude Publica*. 2015;31(2):257-63. DOI:10.1590/0102-311X00041214
5. Capuani L, Bierrenbach AL, Abreu F, Takecian PL, Ferreira JE, Sabino EC. Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. *Cad Saude Publica*. 2014;30(8):1623-32. DOI:10.1590/0102-311X00024914
6. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. New York: Springer; 2012. (Data-centric systems and applications).
7. 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
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 Saude Publica*. 2010;26(7):1431-8. DOI:10.1590/S0102-311X2010000700022
9. Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. *AMIA Annu Symp Proc*. 2003:259-63.
10. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques. New York: Springer Science and Business Media; 2007.

11. Malhão TA , Oliveira GP, Codenotti SB, Moherdaui F. Avaliação da completitude do Sistema de Informação de Agravos de Notificação da Tuberculose, Brasil, 2001-2006. *Epidemiol Serv Saude*. 2014;19(3):245-56. DOI:10.5123/S1679-49742010000300007

12. Migowski A, Chaves RBM, Coeli CM, Ribeiro ALP, Tura BR, Kuschnir MCC, et al. Acurácia do relacionamento probabilístico na avaliação da alta complexidade em cardiologia. *Rev Saude Publica*. 2011;45(2):269-75. DOI:10.1590/S0034-89102011005000012

13. Mohamed GQ, Zhang H. Accuracy of public health data linkages. *Matern Child Health J*. 2009;13(4):531-8. DOI:10.1007/s10995-008-0377-6

14. Moreira CMM, Maciel ELN. Completude dos dados do Programa de Controle da Tuberculose no Sistema de Informação de Agravos de Notificação no Estado do Espírito Santo, Brasil: uma análise do período de 2001 a 2005. *J Bras Pneumol*. 2008;34(4):225-9. DOI:10.1590/S1806-37132008000400007

15. Pacheco AG, Saraceni V, Tuboi SH, Moulton LH, Chaisson RE, Cavalcante SC, et al. Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in Brazil. *Am J Epidemiol*. 2008;168(11):1326-32. DOI:10.1093/aje/kwn249

16. 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

17. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21(20):3940-1. DOI:10.1093/bioinformatics/bti623

18. Van Hest NA, Story A, Grant AD, Antoine D, Croft JP, Watson JM, et al. Record-linkage and capture-recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999-2002. *Epidemiol Infect*. 2008;136(12):1606-16. DOI:10.1017/S0950268808000496

19. World Health Organization. Assessing tuberculosis under-reporting through inventory studies. Geneva: WHO; 2012.

Financiamento: Secretaria de Vigilância em Saúde/Ministério da Saúde (TC 234/12); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – Processos 481654/2012-7, 309728/2012-6 e 305545/2015-9); Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (Faperj – Processo E-26/203.195/2015).

Contribuição dos Autores: Todos os autores participaram de todas as etapas de confecção e aprovação do artigo.

Conflito de Interesses: Os autores declaram não haver conflito de interesses.