CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

Inclusão de etapa de pós-processamento determinístico para o aumento de performance do relacionamento (*linkage*) probabilístico

Inclusion of a deterministic post-processing stage to increase the performance of probabilistic record linkage

Inclusión de la etapa de postprocesamiento determinístico para el aumento del rendimiento del enlace (*linkage*) probabilístico

Rafael Brustulin [1]
Poliana Guerino Marson [2]

doi: 10.1590/0102-311X00088117

Resumo

O objetivo do presente estudo foi demonstrar a aplicação de uma etapa de pós-processamento determinístico, baseada em medidas de similaridade, para aumentar a performance do relacionamento probabilístico com e sem a etapa de revisão manual. As bases de dados utilizadas no estudo foram o Sistema de Informação de Agravos de Notificação e o Sistema de Informações sobre Mortalidade, no período de 2007 a 2015, do Município de Palmas, Tocantins, Brasil. O software probabilístico utilizado foi o OpenRecLink; foi desenvolvida e aplicada uma etapa de pós-processamento determinístico aos dados obtidos por três diferentes estratégias de pareamento probabilístico. As três estratégias foram comparadas entre si e acrescidas da etapa de pós-processamento determinístico. A sensibilidade das estratégias probabilísticas sem revisão manual variou entre 69,1% e 77,8%, já as mesmas estratégias, acrescidas da etapa de pós-processamento determinístico, apresentaram uma variação entre 92,9% e 96,3%. A sensibilidade de duas estratégias probabilísticas com revisão manual foi semelhante à obtida pela etapa de pós-processamento determinístico, no entanto, o número de pares destinados à revisão manual pelas duas estratégias probabilísticas variou entre 1.177 e 1.132 registros, contra 149 e 145 após a etapa de pós-processamento determinístico. Nossos resultados sugerem que a etapa de pós-processamento determinístico é uma opção promissora, tanto para aumentar a sensibilidade quanto para reduzir o número de pares que precisam ser revisados manualmente, ou mesmo para eliminar sua necessidade.

Base de Dados; Software; Processamento Automatizado de Dados; Sistemas de Informação

Correspondência
R. Brustulin
Secretaria Municipal de Saúde de Palmas.
Quadra 804 Sul, Alameda 2, Palmas, TO 77023-028, Brasil.
eurafael@msn.com

[1] Secretaria Municipal de Saúde de Palmas, Palmas, Brasil.
[2] Universidade Federal do Tocantins, Palmas, Brasil.

Introdução

O relacionamento (*linkage*) de bases de dados é um processo que visa a identificar de forma precisa se dois ou mais registros em uma ou mais bases de dados pertencem ao mesmo indivíduo [1]. No Brasil, o seu uso tem aumentado em estudos envolvendo bases de dados de saúde pública, tais como o Sistema de Informação de Agravos de Notificação (SINAN) e o Sistema de Informações sobre Mortalidade (SIM). Nesses estudos, o relacionamento de dados é usado tanto para detectar e remover registros duplicados quanto para integrar informações de bases de dados diferentes [2,3,4,5]. Dessa forma, ajudam a melhorar a qualidade e integridade dos dados, permitem reutilizar dados existentes para responder a uma larga gama de questões, reduzir custos e esforços relacionados à coleta de dados, dentre outros [6,7,8,9].

Os métodos de relacionamento de bases de dados automatizados podem ser divididos em dois grandes grupos: os métodos determinísticos, que empregam conjuntos de regras baseadas em resultados de concordância e discordância entre itens correspondentes; e os métodos probabilísticos, que fazem uso de métodos estatísticos para determinar a concordância ou discordância dos registros, os quais normalmente são classificados mediante a interpretação dos escores gerados pelo método [10]. Vários trabalhos comparam os métodos probabilísticos e determinísticos, em que uma parte dos estudos aponta os métodos determinísticos como os mais acurados [10,11,12,13,14], e outra parte aponta os métodos probabilísticos [15,16,17,18]. Estudos que fazem uso de métodos probabilístico-determinísticos, de forma integrada, ainda são incomuns [19,20].

Devido à versatilidade, os métodos probabilísticos são mais empregados em trabalhos internacionais e nacionais [14,17]. No entanto, o grande número de pares destinados à revisão manual, em estudos nacionais, é o principal problema apresentado pelo método [6,21,22].

Como observado por Pacheco et al. [12], a estratégia mais empregada para aumentar a sensibilidade das técnicas probabilísticas no Brasil é sacrificar a especificidade, resultando num aumento significativo de pares destinados à revisão manual. Para contrapor esse problema, neste estudo propomos demonstrar a aplicação de uma etapa de pós-processamento determinístico (EPPD), baseada em medidas de similaridade, para aumentar a performance do relacionamento probabilístico com e sem a etapa de revisão manual.

Metodologia

Bases de dados e pré-processamento

Foram utilizadas as bases de dados do SINAN e do SIM, no período de 2007 a 2015, do Município de Palmas, Tocantins, disponibilizados pela Secretaria Municipal da Saúde. Quanto ao SINAN, foi utilizada a base completa de notificação individual (SINAN-NET). No estudo, a variável "nome do paciente" para o SINAN e "nome do falecido" para o SIM foi descrita apenas como "nome". O trabalho foi aprovado pelo Comitê de Ética em Pesquisa da Fundação Universidade Federal do Tocantins (parecer nº 1.766.389/2016).

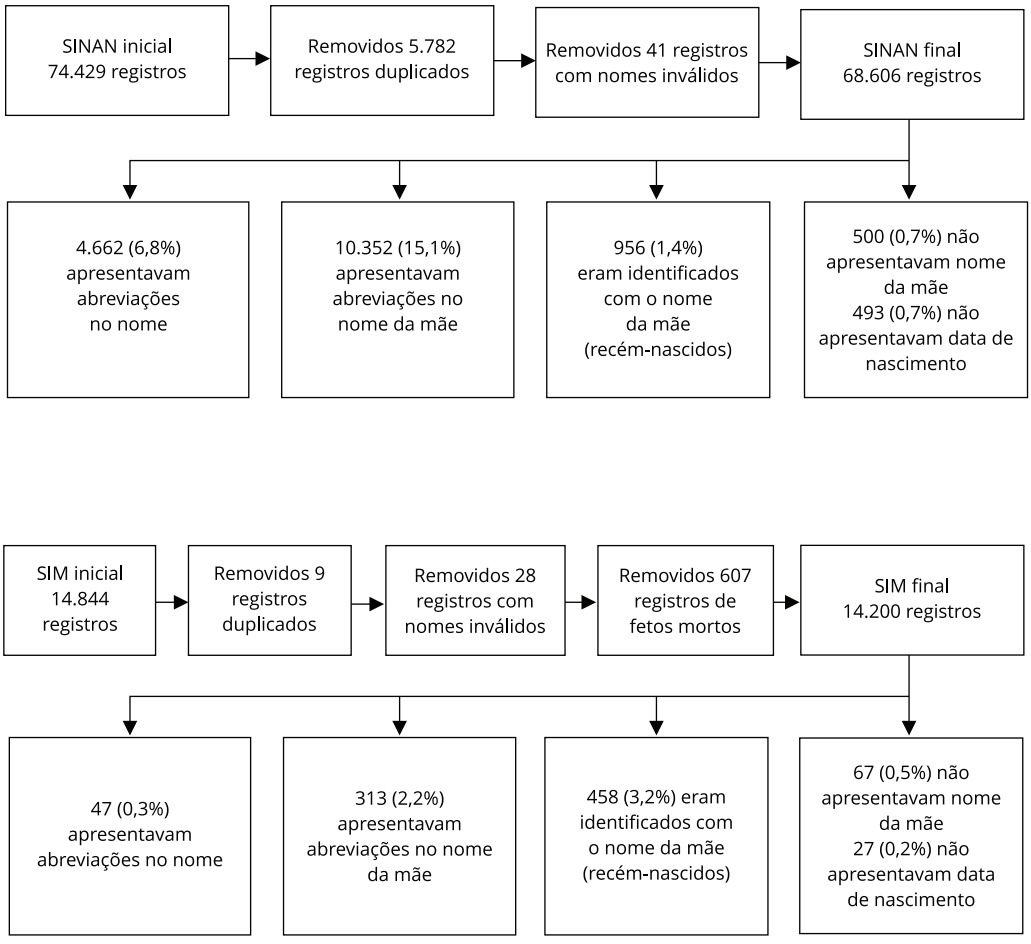
Para as duas bases de dados foram removidas as duplicidades em que o nome e a data de nascimento eram idênticos, isto é, apresentavam as mesmas cadeias de caracteres. Também foram excluídos os registros que apresentavam nomes inválidos. Considerou-se inválidos nomes como: "desconhecido", "indigente", "sem nome", expressos por números (p.ex.: "123456"), dentre outros. Adicionalmente, foram excluídos do SIM os registros relativos a abortos/natimortos, em função da falta de informação sobre a data de nascimento. Por fim, termos como "recém-nascido", "filho de", dentre outros, foram substituídos por "RN" em ambas as bases de dados. Após a etapa de pré-processamento das bases de dados, o SINAN apresentou 68.606 registros e o SIM 14.200 (Figura 1).

Relacionamento probabilístico

O software usado para o relacionamento probabilístico de dados foi o OpenRecLink versão 3.1 (http://reclink.sourceforge.net/). As bases de dados foram importadas e as variáveis foram padroniza-

Figura 1

Pré-processamento e características das bases de dados utilizadas.



SIM: Sistema de Informações sobre Mortalidade; SINAN: Sistema de Informação de Agravos de Notificação.

das utilizando-se as ferramentas do próprio software. Utilizou-se três estratégias de comparação: (1) nome, dia, mês e ano de nascimento; (2) nome, nome da mãe, dia, mês e ano de nascimento; (3) nome, nome da mãe e data de nascimento. Para cada estratégia foram empregados quatro passos de blocagem: (1) *soundex* do primeiro nome e soundex do último nome; (2) *soundex* do primeiro nome, sexo e *soundex* do primeiro nome da mãe; (3) iniciais do nome do meio, sexo e *soundex* do último nome; e (4) data de nascimento. Os nomes foram comparados mediante uso de algoritmos baseados na distância de Levenshtein; e para o nome empregou-se o limiar de 85%; já para o nome da mãe empregou-se 75%. O campo data de nascimento (com 8 dígitos e sem "/") foi avaliado pela utilização do algoritmo que compara caractere a caractere segundo sua posição, empregando-se o limiar de 65%. Já os campos dia, mês e ano de nascimento foram comparados empregando-se um algoritmo para a diferença de valor (+/-2). Os parâmetros para a construção dos fatores de ponderação (concordância ou discordância) foram estimados com base no algoritmo de *expectation-maximization*, usando-se a chave de blocagem formada pela combinação entre o código *soundex* do primeiro e do último nome e o sexo. Todos os algoritmos citados estavam disponíveis no próprio software [23].

Os passos de blocagem foram escolhidos valendo-se da experiência dos autores com bases de dados locais. O número de passos de blocagem foi limitado a quatro porque, do contrário, a revisão manual seria inviabilizada pelo grande número de registros pareados.

Todos os passos de blocagem foram realizados usando-se as bases de dados iniciais, sem remoção dos registros pareados para os passos subsequentes. Os pareamentos formados desse modo não foram revisados manualmente no OpenRecLink. Todos os pareamentos com escore maior que zero foram exportados e consolidados em um único arquivo, para cada estratégia, após a remoção dos pareamentos duplicados formados em passos de blocagem diferentes. As características das estratégias empregadas para o relacionamento probabilístico e número de registros obtidos por estratégia são apresentados na Tabela 1.

O procedimento de exportação dos registros foi adotado porque a etapa de pós-processamento determinístico, que será descrita posteriormente, foi desenvolvida para processar os dados gerados pelo OpenRecLink. Adicionalmente, removendo as duplicidades, os 169.639 registros pareados resultantes das três estratégias somadas (Tabela 1) puderam ser consolidados em um arquivo com 99.588 registros, facilitando a etapa de revisão manual.

Rotina para revisão manual

O arquivo consolidado das três estratégias probabilísticas foi submetido à revisão manual. Foram obedecidos os seguintes critérios para a classificação dos registros pareados como par verdadeiro: (1) "data de nascimento" e "nome" idênticos; (2) "data de nascimento" com erro de um dígito e "nome" idêntico, mas não composto por nomes comuns (p.ex.: João, Maria, José, Ana etc.); (3) "data de nascimento" com erro de até dois dígitos e "nome" idêntico, mas contendo um nome raro; (4) "nome" e "nome da mãe" idênticos, com "data de nascimento" ausente ou contendo erros de até dois dígitos no "dia" e/ou "mês de nascimento" ou no "ano de nascimento"; (5) "nome" incomum e "nome da mãe" idênticos, com "data de nascimento" ausente ou contendo erros expressivos (isto é, erros com mais de dois dígitos); (6) "sexo" e "nome da mãe" idênticos e "data de nascimento" com erro de até um dígito (para registros de recém-nascidos); (7) "sexo" e "nome da mãe" incomum idênticos e "data de nascimento" com erro de até dois dígitos, mas com diferença entre as datas inferior a seis meses (para registros de recém-nascidos).

Na comparação dos registros, o "nome" foi considerado idêntico quando apresentava os mesmos caracteres (letra a letra) ou quando era considerado "igual" pelo revisor, mas apresentava abreviações, omissões ou inversões de parte do nome ou erros tipográficos. Seguiu-se o mesmo critério para o "nome da mãe".

Tabela 1

Características das três estratégias de relacionamento probabilístico.

	Estratégia 1	Estratégia 2	Estratégia 3
Amplitude dos escores	-13,1 a 25,1	-23,5 a 34,9	-9,4 a 33,1
Probabilidades m_i e u_i para o nome	70,9 e 0,001	89,6 e 0,001	94,2 e 0,001
Probabilidades m_i e u_i para o nome da mãe	NA	74,8 e 0,05	79,1 e 0,05
Probabilidades m_i e u_i para a data de nascimento	NA	NA	87,8 e 1,1
Probabilidades m_i e u_i para o dia de nascimento	86,9 e 15,7	91,9 e 15,8	NA
Probabilidades m_i e u_i para o mês de nascimento	99,2 e 37,8	99,9 e 37,9	NA
Probabilidades m_i e u_i para o ano de nascimento	87,4 e 3,9	91,1 e 3,9	NA
Número de pares com escore > 0	89.181	50.382	30.076

NA: não se aplica.

Ao fim do processo, obteve-se um arquivo com 2.005 pares verdadeiros, o qual foi considerado padrão-ouro para a avaliação das estratégias de relacionamento de dados e para a definição dos pontos de corte dos escores do relacionamento probabilístico. Dos 2.005 pares verdadeiros, 90 eram de recém-nascidos.

A estratégia adotada para tornar o processo de revisão mais ágil pode ser consultada no Material Suplementar (cf. http://cadernos.ensp.fiocruz.br/csp/public_site/arquivo/informacao-suplementar_9379.pdf).

Pontos de corte para os escores do relacionamento probabilístico

Foram definidos dois pontos de corte para cada estratégia de pareamento probabilístico com base na análise exploratória das distribuições dos escores e pela inspeção das curvas *precision-recall* (curva PR), usando-se como referência o padrão-ouro. Para o procedimento, foi utilizado o pacote ROCR (versão 1.0-7) [24] do software R (http://www.r-project.org).

O critério de escolha do primeiro ponto de corte (alto) baseou-se em um valor preditivo positivo (VPP) elevado (aproximadamente 98%) e, o segundo ponto de corte (baixo), uma maior sensibilidade, ponderada pelo número de pares destinados à revisão manual.

Etapa de pós-processamento determinístico (EPPD)

A EPPD foi desenvolvida na forma de um suplemento (*add-in*) para o Excel 2013-2016 (Microsoft, Corp., Estados Unidos) por meio do Visual Studio Community 2017, em linguagem VB.NET. O Visual Studio Community 2017 é gratuito para o desenvolvimento de softwares livres (https://www.visualstudio.com/pt-br/downloads/). O suplemento (*add-in*) desenvolvido no estudo pode ser solicitado ao autor principal, mediante pedido por e-mail, além de estar disponível no endereço eletrônico https://1drv.ms/f/s!Ap43rQraSi7EgY9KgvSvc5cNPt2bcQ para *download*.

Em síntese, a proposta do trabalho divide a rotina de relacionamento de dados em duas fases: (fase 1) realizada no OpenRecLink, envolve a etapa de pré-processamento, padronização, blocagem, cálculo dos escores e exportação dos registros pareados; (fase 2) aplicação da EPPD que envolve o emprego das regras determinísticas para a classificação dos registros pareados como "par", "não par" e "duvidoso", revisão manual dos pareamentos duvidosos e consolidação das duas bases analisadas (similar à função do módulo "combina" do OpenRecLink).

Na EPPD, após o arquivo gerado pelo OpenRecLink ser importado, a variável "nome" era dividida em "primeiro nome", "segundo nome" e "último nome". A variável "nome da mãe" era dividida em "primeiro nome da mãe" e "último nome da mãe". Também foi gerada uma tabela de frequência para essas novas variáveis valendo-se da base de dados do SINAN. Foram classificados como "raro" as variáveis com frequência inferior a 0,005%, "não frequente" quando inferior a 0,2%, "frequente" quando inferior a 1% e "muito frequente" quando igual ou superior a 1%. Os limiares foram definidos empiricamente.

Para o cálculo da medida de similaridade das variáveis "primeiro nome", "segundo nome", "último nome", "primeiro nome da mãe" e "último nome da mãe", utilizou-se o mesmo algoritmo empregado pelo OpenRecLink [23,25]. Para as variáveis "nome" e "nome da mãe", a medida de similaridade foi obtida usando-se uma modificação do algoritmo encontrado no OpenRecLink. A modificação do algoritmo proposta reduz o impacto, no valor calculado para as medidas de similaridade, das abreviações e omissões de parte do nome. Para o cálculo da medida de similaridade da "data de nascimento" foi empregado um novo algoritmo desenvolvido para o estudo. A descrição detalhada dos algoritmos alternativos utilizados, bem como o código na linguagem VB.NET, podem ser encontrados no Material Suplementar (cf. http://cadernos.ensp.fiocruz.br/csp/public_site/arquivo/informacao-suplementar_9379.pdf).

As variáveis "primeiro nome", "segundo nome", "último nome", "primeiro nome da mãe" e "último nome da mãe" foram classificadas de acordo com o valor da medida de similaridade. Quando inferior a 0,71 considerou-se "discordante", quando menor que 1, "duvidoso" e, quando igual 1, classificou-se de acordo com a tabela de frequência (p.ex.: "raro", "não frequente", "frequente" e "muito frequente"). O limiar foi definido empiricamente.

Foram utilizadas 36 regras sequenciais para classificar os registros como "par" e "não par". As regras que classificavam os registros como "não par" antecediam as regras que os classificavam como "par"; os registros que não atendiam os critérios das referidas regras eram classificados como "duvidoso". As regras foram construídas baseando-se no conhecimento acumulado na construção do padrão-ouro.

A primeira regra (regra 1 do Quadro 1) teve o objetivo de classificar como "não par" os registros cuja soma da medida de similaridade fosse inferior a 1,7. Os campos "nome da mãe" e/ou "data de nascimento", quando em branco, foram automaticamente pontuados como 0,5 para viabilizar a aplicação da regra. Os valores foram definidos empiricamente com base em simulações de inserções de erros tipográficos.

As demais regras foram divididas em dois grupos: A e B. A divisão dos grupos se deu por meio de restrição, o grupo A continha as regras (10 ao todo) que só eram aplicadas quando detectado o prefixo RN em qualquer um dos nomes dos registros pareados; o grupo B era aplicado quando não era detectado o prefixo RN. Além do prefixo RN, outras condições restringiam o uso de determinadas regras. Alguns exemplos de regras e restrições podem ser vistos no Quadro 1, as regras foram baseadas na discordância e/ou concordância/frequência das partes dos nomes, medida de similaridade das variáveis, do sexo, iniciais do nome e escore probabilístico.

Por exemplo, na segunda regra do Quadro 1, se o registro em questão tiver o "primeiro nome" "raro" ou "não frequente" ou for de um recém-nascido, a regra não é aplicada e o registro é submetido a outra regra. Caso não exista restrição e o registro atenda aos critérios dispostos na regra, o mesmo é classificado como "não par"; se o registro não atender aos critérios, o mesmo é submetido à regra subsequente. Caso o registro não atenda aos critérios de nenhuma regra, o mesmo é classificado como "duvidoso".

Avaliação do relacionamento de dados

As performances das estratégias de relacionamento probabilístico sem a EPPD e acrescidas da EPPD foram comparadas entre si sem a etapa de revisão manual e com a etapa de revisão manual.

Quadro 1

Exemplos de algumas regras utilizadas na etapa de pós-processamento determinístico.

ORDEM	REGRA	RESTRIÇÃO	CLASSIFICAÇÃO
1	SMS < 1,7	NA	Não par
2	Soma da MS do NM e DN < 1,4; Soma da MS do NM e NMM < 1,7; Soma da MS do DN e NMM < 1,9	PN raro ou não frequente; Registro de RN	Não par
3	MS da DN < 0,6; PN é muito frequente ou é discordante; SN é muito frequente ou é discordante, ou não existe; UN é muito frequente ou é discordante	DN não preenchida; PNM é raro; Registro de RN	Não par
4	MS do NMM < 0,8; MS da DN < 0,7	Registro de não RN; PNM raro	Não par
5	SMS > 2,81	PN discordante; Sexo discordante (se SMS < 3); Registro de RN	Par
6	MS da DN > 0,9; MS do NMM > 0,93	Registro de não RN; Sexo discordante	Par

DN: data de nascimento; MS: medida de similaridade; NA: não se aplica; NM: nome; NMM: nome da mãe; PN: primeiro nome; PNM: primeiro nome da mãe; RN: nome identificado com o nome da mãe; SMS: soma da medida de similaridade do nome, nome da mãe e data de nascimento; SN: segundo nome; UN: último nome.

Para a avaliação das estratégias sem a etapa de revisão manual a classificação dos registros pareados ocorreu da seguinte forma: no OpenRecLink considerou-se "par" todos os registros com escores acima do primeiro ponto de corte e, "não par", todos os registros com escores abaixo do primeiro ponto de corte; na EPPD os registros classificados como "duvidoso" pelas regras determinísticas foram reclassificados como "não par".

Para a avaliação das estratégias com a etapa de revisão manual a classificação dos registros pareados ocorreu da seguinte forma: no OpenRecLink utilizaram-se dois pontos de corte para classificar os registros como "par", "duvidoso" (zona cinza) e "não par"; na EPPD os registros pareados foram classificados como "par", "duvidoso" e "não par" de acordo com as regras determinísticas. Antes da avaliação da performance das estratégias, os registros classificados como "duvidoso" (zona cinza) foram reclassificados como "par" e "não par", de acordo com o padrão-ouro, para simular a realização da etapa de revisão manual.

A performance das estratégias de relacionamento de dados, em relação ao padrão-ouro, foi avaliada por meio da comparação de medidas de desempenho. Para a comparação das estratégias sem a etapa de revisão manual foram utilizadas medidas de sensibilidade, especificidade, VPP e *f-measure* [26]. Para a comparação das estratégias com a etapa de revisão manual foram utilizadas medidas de sensibilidade, especificidade, VPP, *f-measure* e número de registros destinados à revisão manual. O intervalo de 95% de confiança (IC95%) foi calculado pelo método exato de Clopper-Pearson usando-se o pacote MKmisc (versão 0.993) no software R.

Resultado

Sem a etapa de revisão manual, as medidas de desempenho das estratégias de relacionamento probabilístico, com e sem a EPPD, são apresentadas na Tabela 2. Sem a EPPD, a estratégia 1 apresentou sensibilidade 11% superior à média das duas outras estratégias. Quando aplicado a EPPD, observaram-se ganhos de performance expressivos para sensibilidade e *f-measure* em todas as estratégias. O maior aumento de sensibilidade ocorreu na estratégia 2, em que a EPPD elevou a sensibilidade em 39,4%.

Com a etapa de revisão manual, as medidas de desempenho das estratégias de relacionamento probabilístico, com ou sem a EPPD, são apresentadas na Tabela 3. Sem a EPPD, a estratégia 1 apresentou sensibilidade 11,5% inferior à média das duas outras estratégias. Embora a EPPD resultasse em aumento de sensibilidade em todas as estratégias, o aumento foi pouco expressivo, sendo o menor aumento de 1,9% na estratégia 3. Por outro lado, o acréscimo da EPPD resultou em significativa redução no número de pares destinados à revisão manual. Na estratégia 2, que apresentou a maior sensibilidade, a redução no número de pares destinados à revisão manual foi de 87,3%.

A estratégia 3 seguida da EPPD, com ou sem a etapa de revisão manual, apresentou sensibilidade inferior a 3% em relação à média das duas outras estratégias (Tabelas 2 e 3). Isso aconteceu porque 78 dos 2.005 pares verdadeiros apresentaram escores inferiores a zero.

A estratégia 1 seguida da EPPD, com ou sem a etapa de revisão manual, considerando-se o IC95%, apresentou sensibilidade semelhante à estratégia 2 (Tabelas 2 e 3), porém, a estratégia 1 apresentou 77% mais pares com escores superiores a zero em relação à estratégia 2 (Tabela 1). Dessa forma, a estratégia 2 seguida da EPPD apresentou resultados mais satisfatórios, tanto para as rotinas sem a etapa de revisão manual quanto para as com a etapa de revisão manual. É válido destacar que a estratégia 2 seguida da EPPD sem revisão manual apresentou sensibilidade 2,1% maior que a estratégia 2 sem a EPPD seguida de revisão manual.

A regra zero foi a regra determinística mais eficiente para a classificação dos registros pareados. Sozinha, foi capaz de classificar corretamente como "não par", na estratégia 2, 73,2% (36.898/50.376) dos registros pareados com escores superiores a zero.

Dos 1.177 registros pareados dentro da zona cinza, na estratégia 2 sem a EPPD (Tabela 3), 670 (56,9%) eram pares falsos e 507 (43,1%) eram pares verdadeiros, segundo o padrão-ouro. Após a aplicação da EPPD, apenas 60 (5,1%) registros permaneceram classificados como "duvidoso", 648 (55,1%) foram classificados como "não par" e 469 (39,8%) como "par".

Dos 670 pares falsos, dentro da zona cinza na estratégia 2 sem a EPPD (Tabela 3), 408 (60,9%) pertenciam a um grupo de registros que apresentavam nomes muito similares e comuns (p.ex.: José

Tabela 2

Performance das estratégias de relacionamento de dados sem a etapa de revisão manual, em relação ao padrão-ouro.

	Estratégia 1		Estratégia 2		Estratégia 3	
	Pb	Pb+EPPD	Pb	Pb+EPPD	Pb	Pb+EPPD
PC	20	NA	25	NA	23,5	NA
VP	1.560	1.927	1.385	1.931	1.425	1.863
FP	30	0	27	0	37	0
VN	66.571	66.601	66.574	66.601	66.564	66.601
FN	445	78	620	74	580	142
Sensibilidade (IC95%)	77,8 (75,9-79,6)	96,1 (95,2-96,9)	69,1 (67,0-71,1)	96,3 (95,4-97,1)	71,1 (69,0-73,0)	92,9 (91,7-94,0)
Especificidade (IC95%)	99,9 (99,9-100,0)	100,0 (99,9-100,0)	99,9 (99,9-100,0)	100,0 (99,9-100,0)	99,9 (99,9-100,0)	100,0 (99,9-100,0)
f-measure (IC95%)	86,8 (85,6-87,9)	98,0 (97,50-98,4)	81,1 (79,7-82,4)	98,1 (97,6-98,5)	82,2 (80,9-83,5)	96,3 (95,7-96,9)
VPP (IC95%)	98,1 (97,3-98,7)	100,0 (99,8-100,0)	98,1 (97,2-98,7)	100,0 (99,8-100,0)	97,5 (96,5-98,2)	100,0 (99,8-100,0)

IC95%: intervalo de 95% de confiança; FN: falso-negativo; FP: falso-positivo; NA: não se aplica; Pb: método probabilístico; Pb+EPPD: método probabilístico seguido de etapa de pós-processamento determinístico; PC: ponto de corte; VN: verdadeiro-negativo; VP: verdadeiro-positivo; VPP: valor preditivo positivo.

Tabela 3

Performance das estratégias de relacionamento de dados com a etapa de revisão manual, em relação ao padrão-ouro.

	Estratégia 1		Estratégia 2		Estratégia 3	
	Pb	Pb+EPPD	Pb	Pb+EPPD	Pb	Pb+EPPD
PC1	20	NA	25	NA	23,5	NA
PC2	15,5	NA	13	NA	11	NA
VP	1.670	1.992	1.892	2.000	1.886	1.923
FP	30	0	27	0	37	0
VN	66.571	66.601	66.574	66.601	66.564	66.601
FN	335	13	113	5	119	82
Sensibilidade (IC95%)	83,3 (81,6-84,9)	99,4 (98,9-99,6)	94,4 (93,3-95,3)	99,8 (99,4-99,9)	94,1 (92,9-95,1)	95,9 (94,9-96,7)
Especificidade (IC95%)	99,9 (99,9-99,9)	100,0 (99,9-100,0)	99,9 (99,9-100,0)	100,0 (99,9-100,0)	99,9 (99,9-100,0)	100,0 (99,9-100,0)
f-measure (IC95%)	90,1 (89,1-91,1)	99,7 (99,4-99,8)	96,4 (95,8-97,0)	99,9 (99,7-99,9)	96,0 (95,4-96,6)	97,9 (99,4-98,3)
VPP IC95%	98,2 (97,5-98,8)	100,0 (99,8-100,0)	98,6 (98,0-99,1)	100,0 (99,8-100,0)	98,1 (97,4-98,6)	100,0 (99,8-100,0)
RM	568	150	1.177	149	1.132	145

IC95%: intervalo de 95% de confiança; FN: falso-negativo; FP: falso-positivo; NA: não se aplica; Pb: método probabilístico; Pb+EPPD: método probabilístico seguido de etapa de pós-processamento determinístico; PC1: ponto de corte alto; PC2: ponto de corte baixo; RM: número de pares destinados à revisão manual; VN: verdadeiro-negativo; VP: verdadeiro-positivo; VPP: valor preditivo positivo.

Pereira Silva), com datas de nascimentos com erros de 2 ou mais dígitos. Já para os 507 pares verdadeiros, 208 (41%) pertenciam a um grupo de registros que registravam erros isolados, tais como: 163 (32,1%) apresentavam abreviações ou omissões no "nome" e/ou "nome da mãe", 29 (5,7%) não apresentavam o "nome da mãe", 14 (2,8%) apresentavam nomes diferentes para o "nome da mãe" e 2

(0,4%) não registravam a "data de nascimento". Quando a EPPD era acrescida, ambos os grupos eram classificados como "não par" e "par", respectivamente. Os demais registros eram bem heterogêneos e apresentavam uma combinação de erros (por exemplo, erros tipográficos no "nome" e abreviação do "nome da mãe").

Dos 113 falso-negativos, da estratégia 2 sem a EPPD (Tabela 3), 51 (45,1%) continham abreviações/omissões no "nome" e "nome da mãe", "data de nascimento" idênticas e apresentavam escores inferiores a 3,3. Esses mesmos registros foram classificados como "par" quando aplicada a EPPD, e eram mais numerosos do que os 14 (12,4%) pares verdadeiros sem "data de nascimento" e 3 (2,6%) pares verdadeiros sem o "nome da mãe" encontrados nessa região.

A influência dos registros de recém-nascidos para os resultados da estratégia 2, sem a EPPD, foi pequena, apenas 77 (6,5%) registros que pertenciam a este grupo e encontravam-se na zona cinza, sendo 43 (3,6%) pares falsos, já abaixo da zona cinza constavam apenas 4 dos 90 pares verdadeiros.

Discussão

Dentre os principais achados do estudo destacam-se: (1) a EPPD foi capaz de aumentar a performance do método probabilístico em todos os cenários avaliados; (2) o número de pares destinados à revisão manual foi substancialmente reduzido quando aplicada a EPPD; (3) a estratégia 2 seguida da EPPD foi mais eficiente tanto na rotina sem a etapa de revisão manual quanto na rotina com a etapa de revisão manual; e (4) a estratégia 2 com a EPPD e sem a etapa de revisão manual foi mais sensível e específica do que a mesma estratégia sem a EPPD e com a etapa de revisão manual.

O ganho de performance das medidas de desempenho em todas as estratégias quando acrescidas da EPPD já era esperado, uma vez que se trata de um acréscimo metodológico para contrapor as deficiências do OpenRecLink, que tende a apresentar um número elevado de falso-positivos com escores intermediários [11,13] e um pequeno, mas relevante, número de falso-negativos com escores baixos [27].

O OpenRecLink calcula os escores dos registros pareados sem levar em consideração a tabela de frequência dos nomes, resultando em um número considerável de pares falsos com escores intermediários [16]. No presente trabalho, essa deficiência foi responsável por formar um grupo de registro que correspondeu a mais de 60% dos pares falsos encontrados na zona cinza da estratégia 2, sem a EPPD. Nossos dados apontam que essa deficiência pode ser corrigida aplicando-se a EPPD, que contém regras que levam em consideração a tabela de frequência dos nomes. O uso de tabela de frequência dos nomes é relatado como relevante para o aumento da especificidade tanto em estudos empregando métodos probabilísticos quanto determinísticos [14,16,19,28].

Segundo Oliveira et al. [14], um fator relevante para especificidade e sensibilidade do relacionamento de dados são nomes com medidas de similaridade inferiores a 0,7; medidas de similaridade entre 0,5 e 0,7 são críticas por abrangerem tanto nomes que não pertencem à mesma pessoa quanto nomes de um mesmo indivíduo contendo abreviações ou omissões [29]. É notória a proporção de registros com nomes abreviados no SINAN, 1 a cada 14, bem como a diferença em relação ao SIM, 1 a cada 302, que foi encontrada no estudo (Figura 1). Embora as abreviações de nomes fossem descritas em outros estudos como erro de preenchimento que dificulta o processo de relacionamento de dados [27,30,31], a frequência de nomes com abreviação não foi informada.

Indiretamente, Migowski et al. [27] relataram que 2% dos 1.411 pares verdadeiros encontrados apresentavam abreviações/omissões no nome, escores baixos e sem apresentar erros na data de nascimento; valor similar aos 2,5% dos pares verdadeiros (51/2.005) encontrados no atual estudo, sugerindo que a frequência de nomes com abreviações/omissões sejam similares em ambos os estudos. Nossos resultados demonstram que um terço dos pares verdadeiros com escores intermediários e baixos, na estratégia 2 sem a EPPD, apresentavam apenas a abreviação/omissão de nomes como erro. Esse mesmo grupo de registros pode ser classificado como "par", sem a necessidade de revisão manual, caso a EPPD seja aplicada.

Por último, está um grupo de registros de pares verdadeiros com valores faltantes para variáveis-chave e que apresentam escores intermediários ou baixos [5,10,18,32]. No presente trabalho, a ocorrência de registros desse grupo foi menos frequente em relação aos dois grupos citados anteriormente.

Contudo, a EPPD apresentou regras que previam variáveis faltantes, contribuindo para o aumento da sensibilidade e redução no número de pares destinados à revisão manual apresentado pelo estudo.

A EPPD pode não apresentar ganhos de performance que justifiquem seu emprego em bases de dados de pequeno porte e/ou com alta qualidade de preenchimentos das variáveis-chave, porém estas características podem não ser frequentes em bases de dados de saúde no país [1,33,34,35].

Não é incomum, na literatura, estudos que revisam manualmente registros pareados com escores próximos a zero [2,27,36], ou mesmo inferiores a zero [37]. Se tal medida fosse adotada, a diferença no número de pares destinados à revisão manual entre as estratégias probabilistas sem e com a EPPD seria substancialmente maior do que a reportada pelo estudo. Contudo, trabalhos que buscam alternativas viáveis para eliminar a etapa de revisão manual são raros no âmbito nacional [14,16], contrastando com o aumento do tamanho das bases de dados, que torna o emprego da revisão manual cada vez mais proibitivo [10,27], assim o conceito de EPPD apresentado pelo estudo pode vir a ser uma opção a mais a se considerar no planejamento de projetos de relacionamento de dados.

Um dos poucos exemplos de método probabilístico-determinístico disponíveis é o software Link King que utiliza o método determinístico para identificar registros de possíveis gêmeos e registros em que são possíveis pares verdadeiros pelo método probabilístico, mas que foram classificados como "não par" pelo método determinístico [19]. Assim, o Link King utiliza o método determinístico para aumentar a especificidade do relacionamento de dados, ampliando o número de pares destinados à revisão manual, diferindo da proposta do atual estudo que visa à redução no número de pares destinados à revisão manual.

É importante ressaltar que os valores de sensibilidade, especificidade e VPP foram obtidos para comparação das diferentes estratégias e estão limitados à forma como o padrão-ouro foi obtido. A escolha dos passos de blocagem e não considerar escores negativos eventualmente resultaram em perdas de pares verdadeiros, o que impede que comparações sejam feitas com outros estudos como o de Coutinho & Coeli [37], em que a sensibilidade e especificidade foram calculadas usando-se, como padrão-ouro, informações obtidas por meio do seguimento ativo de uma coorte.

Duas limitações importantes do trabalho são a utilização de bases de dados de pequeno porte e regionalizadas (esfera municipal) e empregar apenas um revisor para a etapa de revisão manual. Essas limitações podem se tornar evidentes quando a EPPD, desenvolvida no estudo, for aplicada, por outros pesquisadores, a outras bases de dados. Isso é esperado porque as regras determinísticas foram construídas utilizando-se limiares definidos empiricamente para análise das bases de dados usadas no estudo. Portanto, espera-se que outras bases de dados apresentem determinadas situações não previstas na construção e definição dos limiares das regras, além das eventuais divergências causadas pela subjetividade da etapa de revisão manual.

Para contornar tais limitações, futuros estudos são necessários para o aprimoramento das regras utilizadas e determinar quais podem ser aplicadas em diferentes bases de dados e quais são restritas à região geográfica ou a determinadas bases de dados.

Conclusão

O presente trabalho contribuiu ao demonstrar a implementação de uma etapa determinística, após o relacionamento de dados obtidos pelo OpenRecLink, como uma opção promissora, tanto para aumentar sensibilidade quanto para reduzir o número de pares que precisam ser revisados manualmente, ou mesmo para eliminar a etapa de revisão manual, quando a mesma for inviável.

Por se tratar de uma estratégia de pós-processamento, o resultado final obtido depende da estratégia de pareamento e blocagem escolhida para o relacionamento probabilístico, além das regras determinísticas empregadas. Mais estudos são necessários para avaliar a robustez da EPPD em diferentes bases de dados, principalmente as de grande porte.

Colaboradores

R. Brustulin participou da concepção e projeto do estudo, interpretação dos dados e redação do artigo. P. G. Marson participou da interpretação dos dados, revisão crítica do conteúdo intelectual e aprovação final da versão a ser publicada.

Agradecimentos

Cláudia T. Fulanetto Costa, pela colaboração preliminar na revisão ortográfica e editoração do artigo.

Referências

1. 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 Saúde Colet (Rio J.) 2006; 14:197-224.
2. Bartholomay P, Oliveira GP, Pinheiro RS, Vasconcelos AMN. Melhoria da qualidade das informações sobre tuberculose a partir do relacionamento entre bases de dados. Cad Saúde Pública 2014; 30:2459-70.
3. Soeiro CMO, Miranda AE, Saraceni V, Santos MC, Talhari S, Ferreira LCL. Syphilis in pregnancy and congenital syphilis in Amazonas State, Brazil: an evaluation using database linkage. Cad Saúde Pública 2014; 30:715-23.
4. Rossetto EV, Luna EJA. Relacionamento entre bases de dados para vigilância da pandemia de influenza A(H1N1)pdm09, Brasil, 2009-2010. Cad Saúde Pública 2016; 32:e00014115.
5. Paixão ES, Harron K, Andrade K, Teixeira MG, Fiaccone RL, Costa MCN, et al. Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in Brazil. BMC Med Inform Decis Mak 2017; 17:108.
6. 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 Saúde Pública 2014; 30:1623-32.
7. Oliveira PPV, Azevedo e Silva G, Curado MP, Malta DC, Moura L. Confiabilidade da causa básica de óbito por câncer entre Sistema de Informações sobre Mortalidade do Brasil e Registro de Câncer de Base Populacional de Goiânia, Goiás, Brasil. Cad Saúde Pública 2014; 30:296-304.
8. Rocha MS, Oliveira GP, Aguiar FP, Saraceni V, Pinheiro RS. Do que morrem os pacientes com tuberculose: causas múltiplas de morte de uma coorte de casos notificados e uma proposta de investigação de causas presumíveis. Cad Saúde Pública 2015; 31:709-21.
9. Spineti PPM, Souza AS, Feijó LA, Garcia MI, Xavier SS. Acurácia do relacionamento probabilístico de registros na identificação de óbitos em uma coorte de pacientes com insuficiência cardíaca descompensada. Cad Saúde Pública 2016; 32:e00097415.
10. Joffe E, Byrne MJ, Reeder P, Herskovic JR, Johnson CW, McCoy AB, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. J Am Med Inform Assoc 2014; 21:97-104.
11. Coutinho R, Coeli C. Sensibilidade do linkage probabilístico na identificação de nascimentos informados: Estudo Pró-Saúde. Rev Saúde Pública 2008; 42:1097-100.

12. 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; 168:1326-32.
13. 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.
14. Oliveira GP, Bierrenbach ALS, Camargo Júnior KR, Coeli CM, Pinheiro RS. Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis. Rev Saúde Pública 2016; 50:1-10.
15. Roos L, Walld R, Wajda A, Bond R, Hartford K. Record linkage strategies, outpatient procedures, and administrative data. Med Care 1996; 34:570-82.
16. Queiroz OV, Guerra Júnior AA, Machado CJ, Andrade ELG, Meira Júnior W, Acúrcio FA, et al. A construção da Base Nacional de Dados em Terapia Renal Substitutiva (TRS) centrada no indivíduo: relacionamento dos registros de óbitos pelo Subsistema de Autorização de Procedimentos de Alta Complexidade (APAC/SIA/SUS) e pelo Sistema de Informações sobre Mortalidade (SIM) – Brasil, 2000-2004. Epidemiol Serv Saúde 2009; 18:107-20.
17. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011; 64:565-72.
18. Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S. When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform 2015; 56:80-6.
19. Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. Health Informatics J 2008; 14:5-15.
20. Cherchiglia ML, Giordano LFC, Machado EL, Gomes IC, Carmo RA, Acúrcio FA, et al. Incidência de hepatite viral C em pacientes em hemodiálise no Brasil entre 2000 e 2003. Cad Saúde Pública 2016; 32:e00167914.
21. Camargo Jr. KR, Coeli CM. RecLink 3: nova versão do programa que implementa a técnica de associação probabilística de registros (probabilistic record linkage). Cad Saúde Colet (Rio J.) 2006; 14:399-404.
22. Coeli CM, Pinheiro RS, Camargo Jr. KR. Conquistas e desafios para o emprego das técnicas de record linkage na pesquisa e avaliação em saúde no Brasil. Epidemiol Serv Saúde 2015; 24:795-802.
23. Camargo Jr. KR, Coeli CM. Going open source: some lessons learned from the development of OpenRecLink. Cad Saúde Pública 2015; 31:257-63.
24. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005; 21:3940-1.
25. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Heidelberg: Springer; 2012.
26. Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication. In: Guillet F, Hamilton HJ, editors. Quality measures in data mining. Heidelberg: Springer; 2007. p. 127-51.
27. 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 Saúde Pública 2011; 45:269-75.
28. Zhu VJ, Overhage MJ, Egg J, Downs SM, Grannis SJ. An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling. J Am Med Inform Assoc 2009; 16:738-45.
29. Snae C. A comparison and analysis of name matching algorithms. Int J Appl Sci Eng Technol 2007; 21:252-7.
30. Peres SV, Latorre MRDO, Michels FAS, Tanaka LF, Coeli CM, Almeida MF. 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.) 2014; 22:428-36.
31. 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.
32. Ansolabehere S, Hersh ED. ADGN: an algorithm for record linkage using address, date of birth, gender, and name. Stat Public Policy (Phila) 2017; 4:1-10.
33. Coeli CM, Barbosa FS, Brito AS, Pinheiro RS, Camargo Jr. KR, Medronho RA, et al. Estimativas de parâmetros no *linkage* entre os bancos de mortalidade e de hospitalização, segundo a qualidade do registro da causa básica do óbito. Cad Saúde Pública 2011; 27:1654-8.
34. Pinto IV, Ramos DN, Esteves C, Belo C, Ferreira T, Rebelo MS. Completude e consistência dos dados dos registros hospitalares de câncer no Brasil. Cad Saúde Colet (Rio J.) 2012; 20:113-20.
35. Teixeira CLS, Bloch KV, Klein CH, Coeli CM. Método de relacionamento de bancos de dados do Sistema de Informações sobre Mortalidade (SIM) e das autorizações de internação hospitalar (BDAIH) no Sistema Único de Saúde (SUS), na investigação de óbitos de causa mal-definida no Estado do Rio de Janeiro. Epidemiol Serv Saúde 2006; 15:47-57.
36. Sousa MH, Cecatti JG, Hardy E, Serruya SJ. Relacionamento probabilístico de registros: uma aplicação na área de morbidade materna grave (*near miss*) e mortalidade materna. Cad Saúde Pública 2008; 24:653-62.
37. Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida. Cad Saúde Pública 2006; 22:2249-52.

Abstract

The aim of this study was to demonstrate the application of a deterministic post-processing stage, based on measures of similarity, to increase the performance of probabilistic record linkage with and without manual revision. The databases used in the study were the Brazilian Information System for Notificable Diseases and the Brazilian Mortality Information System, from 2007 to 2015, in Palmas, Tocantins State, Brazil. The probabilistic software was OpenRecLink, and a deterministic post-processing stage was applied to the data obtained from three different probabilistic linkage strategies. The three strategies were compared to each other, and the deterministic post-processing stage was added. The sensibility of the probabilistic strategies without manual revision varied from 69.1% and 77.8%, while the same strategies plus the deterministic post-processing stage varied from 92.9% to 96.3%. Sensitivity of the two probabilistic strategies with manual revision was similar to that obtained by the deterministic post-processing stage, but the number of matches that were referred to manual revision by the two probabilistic strategies varied between 1,177 and 1,132 records, compared to 149 and 145 after the deterministic post-processing stage. Our findings suggest that the deterministic post-processing stage is a promising option, both to increase the sensitivity and to reduce the number of matches that need to be reviewed manually, or even to eliminate the need for manual revision altogether.

Database; Software; Automatic Data Processing; Information Systems

Resumen

El objetivo del presente estudio fue mostrar la aplicación de una etapa de postprocesamiento determinístico, basada en medidas de similitud, con el objeto de aumentar el rendimiento del enlace probabilístico con y sin etapa de revisión manual. Las bases de datos utilizadas en el estudio fueron el Sistema de Información sobre Enfermedades de Notificación Obligatoria y el Sistema de Informaciones sobre Mortalidad, durante el período de 2007 a 2015, en el municipio de Palmas, Tocantins, Brasil. El software probabilístico utilizado fue el OpenRecLink; se desarrolló y aplicó una etapa de postprocesamiento determinístico con los datos obtenidos mediante tres estrategias diferentes de emparejamiento probabilístico. Las tres estrategias se compararon entre sí y se añadieron a la etapa de postprocesamiento determinístico. La sensibilidad de las estrategias probabilísticas sin revisión manual varió entre el 69,1% y el 77,8%, incluso las mismas estrategias, añadidas de la etapa de postprocesamiento determinístico, presentaron una variación entre 92,9% y 96,3%. La sensibilidad de las dos estrategias probabilísticas con revisión manual fue semejante a la obtenida por la etapa de postprocesamiento determinístico, sin embargo, el número de pares destinados a la revisión manual por las dos estrategias probabilísticas varió entre 1.177 y 1.132 registros, frente 149 y 145 tras la etapa de postprocesamiento determinístico. Nuestros resultados sugieren que la etapa de postprocesamiento determinístico es una opción prometedora, tanto para aumentar la sensibilidad, como para reducir el número de pares que necesitan ser revisados manualmente, o incluso para eliminar su necesidad.

Base de Datos; Programas Informáticos; Procesamiento Automatizado de Datos; Sistemas de Información

Recebido em 23/Mai/2017
Versão final reapresentada em 24/Jan/2018
Aprovado em 12/Mar/2018