DOI: 10.1590/1414-462X201400040017

Artigo Original

Determinação de um ponto de corte para a identificação de pares verdadeiros pelo método probabilístico de *linkage* de base de dados

Determining a cutoff point for identifying the true pairs probabilistic record linkage database

Stela Verzinhasse Peres[1], Maria do Rosário Dias de Oliveira Latorre[2], Fernanda Alessandra Silva Michels[3], Luana Fiengo Tanaka[4], Claudia Medina Coeli[5], Márcia Furquim de Almeida[6]

Resumo

O objetivo deste estudo foi propor pontos de corte nos escores calculados no processo de *linkage* probabilístico, para as diversas topografias de câncer. Neste estudo foi utilizada a base de dados do RCBP-SP, composta por 343.306 casos incidentes de câncer do município de São Paulo, registrados no período de 1997 a 2005, com idades que variaram de menos um a 106 anos, de ambos os sexos. Para o ***linkage*** probabilístico, realizado no programa Reclink III, foram utilizadas a base de dados do PRO-AIM e APAC-SIA/SUS. Foram calculados os valores da área sob a curva, sensibilidade e especificidade para determinar o ponto de corte do escore de maior precisão na identificação dos pares verdadeiros. Na análise das topografias, verificou-se que o ponto de corte no escore 18 apresentou boa acurácia, com valores de sensibilidade que variaram de 73,7 a 96,7% e de especificidade de 98,5 a 99,4%. Conclui-se que, acima do escore 18 encontravam-se quase a totalidade dos pares verdadeiros, enquanto que abaixo deste, menos de 1% dos registros vinculados eram verdadeiros.

Palavras-chave: sistemas de informação; registro médico coordenado; neoplasias.

Abstract

The aim of this study was to propose cut-off points for scores calculated in the probabilistic record linkage process for several cancer topographies. In this study we used the PBCR-SP database composed of 343,306 incident cancer cases from the municipality of São Paulo, registered from 1997 through 2005, aged from less than one to 106 years, of both sexes. PRO-AIM and APAC-SIA/SUS databases were used to probabilistic record linkage using Reclink III software. Area under the curve, sensitivity and specificity values were calculated to determine the cut-off point with the highest accuracy in identifying true matches. In the topography analyses, it was found that the cut-off at score 18 showed good accuracy, with sensitivity ranging from 73.7 to 96.7% and specificity ranging from 98.5 to 99.4%. We concluded that above score 18 nearly all true pairs were found. Whereas, below this cut-off, less than 1% of linked records were true matches.

Keywords: information systems; medical record linkage; neoplasms.

Trabalho realizado na Registro de Câncer de Base Populacional do Município de São Paulo – São Paulo (SP), Brasil.
[1]Doutor em Ciências pela Faculdade de Saúde Pública da Universidade de São Paulo (USP); Analista de Sistemas de Informação em Saúde na Faculdade de Saúde Pública – São Paulo (SP), Brasil.
[2]Doutora em Saúde Pública pela Universidade de São Paulo (USP); Professora Titular do Departamento de Epidemiologia na Universidade de São Paulo (USP) – São Paulo (SP), Brasil.
[3]Doutora em Ciências pela Faculdade de Saúde Pública da Universidade de São Paulo; Aluna de pós-doutorado na Harvard University – Cambridge, Estados Unidos.
[4]Mestre em Ciências pela Faculdade de Saúde Pública da Universidade de São Paulo; Doutoranda em Ciências da Saúde na USP – São Paulo (SP), Brasil.
[5]Doutora em Saúde Coletiva pela Universidade do Estado do Rio de Janeiro; Professora Adjunta no Instituto de Estudos de Saúde Coletiva da Universidade Federal do Rio de Janeiro (IESC/UFRJ) – Rio de Janeiro (RJ), Brasil.
[6]Doutora em Saúde Pública pela Universidade de São Paulo; Professora Associada na Faculdade de Saúde Pública da USP – São Paulo (SP), Brasil.
Endereço para correspondência: Stela Verzinhasse Peres – Rua Benjamim de Laborde,131 – Jardim São Ricardo – CEP: 05143-140 – São Paulo (SP), Brasil – E-mail: svperes@usp.br
Fonte de financiamento: Registro de Câncer de Base Populacional do Município de São Paulo.
Conflito de interesses: nada a declarar.

INTRODUÇÃO

A disponibilidade de grandes bases de dados informatizadas em saúde tornou a técnica de relacionamento de bases de dados, também conhecida como *linkage,* uma alternativa para diferentes tipos de estudos. Esta técnica pode proporcionar a geração de uma base de dados mais completa e de baixo custo operacional[1-4]. No processo de *linkage*, dois métodos são utilizados: o determinístico ou lógico e o probabilístico.

O *linkage* probabilístico é capaz de identificar um indivíduo em diferentes bases de dados mesmo que estes não possuam identificadores unívocos comuns e em bases que apresentem problemas de inconsistência, erros e informações não declaradas. Todavia, nesta técnica, o poder de discriminação das variáveis utilizadas no processo de relacionamento é influenciado pelo número de valores válidos e suas uniformidade. Em relacionamentos de bases de dados que utilizam poucos campos e os mesmos apresentam baixo poder discriminatório, os erros de homônimos tendem a aumentar o número de falsos-positivos[5-9].

Nos últimos anos, foram desenvolvidas pesquisas para avaliar a acurácia do *linkage* probabilístico, na qual esta técnica faz uso de um escore que verifica o quão verossimilhante os registros pareados pertencem à mesma pessoa; isto é, dentre os registros, quais formam pares verdadeiros. Em países como Holanda, Áustria, Escócia e França estudos com diferentes tipos de bases de dados mostraram uma precisão de 89,0%, valores de sensibilidade e especificidade de 97,0 e 99,0%, respectivamente[10-14]. No Brasil, algumas pesquisas também avaliaram a acurácia do método probabilístico em identificar um par verdadeiro entre os registros pareados. Nestas, os valores de sensibilidade variaram de 85,5 a 90,6%, e a especificidade de 99,0 a 100%[15-17].

Contudo, a técnica de *linkage* probabilístico em bases de dados de grande porte, como os Registros de Câncer de Base Populacional (RCBP), ainda é pouco utilizada no Brasil. Este fato está relacionado ao grau de incerteza que existe nos escores gerados e ao número de registros pareados no processo, dado que o *linkage* é o produto da multiplicação dos registros entre duas bases. Assim, a identificação de um par verdadeiro somente pelo maior escore poderia acarretar na perda de informações relevantes, sendo indispensável uma revisão manual extensa, tornando-se inviável na rotina de serviços de Sistemas de Informação.

Neste contexto e sabendo da aplicabilidade da técnica de *linkage* na melhora da qualidade do dado e na completitude da informação, o objetivo do estudo foi propor um ponto de corte, a partir dos escores calculados no processo de *linkage* probabilístico, para que seja possível identificar o maior número de pares verdadeiros com o máximo de acurácia, viabilizando uma rotina segura e rápida para a atualização da base dos dados do RCBP do município de São Paulo (RCBP-SP). As bases de dados utilizadas foram a do RCBP-SP, para as diversas topografias de câncer, *versus* a base do Programa de Aprimoramento de Mortalidade (PRO-AIM) e Autorização e Procedimentos de Alta Complexidade do Sistema de Informações Ambulatoriais do Sistema Único de Saúde (APAC-SIA/SUS).

MATERIAL E MÉTODOS

O estudo da base de dados do RCBP-SP compreendeu o período disponível entre os anos de 1997 a 2005, composta por 343.306 casos incidentes de câncer incluindo tumores *in situ* (C00.0 a C80.9) do município de São Paulo, com idades que variam de menores de 1 a 106 anos, de ambos os sexos.

A base de dados do PRO-AIM apresentou 767.752 óbitos, exceto os fetais, ocorridos entre 1997 a 2007, no município de São Paulo. A escolha de dois anos além do seguimento do RCBP-SP foi para captar um número maior de óbitos, além de este ser o período disponibilizado pela Coordenação de Epidemiologia e Informação (CEINFO).

A terceira base de dados foi a APAC-SIA/SUS referente a 863.735 pacientes, de todas as idades e de ambos os sexos, que realizaram qualquer tratamento/procedimento entre o período de agosto de 2003 a dezembro de 2007 no município de São Paulo. A escolha se deve à necessidade de identificar os pacientes vivos, pois, nesta, encontram-se aqueles que receberam ou realizaram qualquer procedimento de alta complexidade e custo. Foram solicitados à Secretaria Estadual de Saúde de São Paulo (SES-SP) os dados referentes as APAC-SIA/SUS para o período de 1997 a 2007/ porém, só foi possível obter as informações do período de agosto de 2003 a dezembro de 2007, disponibilizadas pela Secretaria Municipal de Saúde de São Paulo (SMS-SP).

Para a realização do *linkage* probabilístico foram seguidas as etapas de padronização, blocagem e pareamento dos registros[18,19], realizadas pelo programa Reclink 3.1.6[20].

PREPARAÇÃO DAS BASES DE DADOS

A primeira etapa deste processo foi a limpeza das bases do PRO-AIM e APAC-SIA/SUS. Para o PRO-AIM, foram excluídos 24.758 óbitos fetais e 6.542 óbitos que estavam sem informação ou apresentavam a palavra "indigente" no campo "nome". Para a base de dados da APAC-SIA/SUS, que, originalmente, continha 31.743.533 registros, realizou-se o processo de identificação de pacientes que apareceram mais de uma vez (dado que, para o *linkage* probabilístico, esta base de dados deveria apresentar um único registro para cada paciente e, neste, estar presente a data da última vez em que um procedimento ou medicamento foi

solicitado). Para identificação, foi considerada a igualdade das variáveis "nome do paciente", "data de nascimento", "Cadastro de Pessoa Física (CPF)" (para os que possuíam esta informação) e "nome da mãe". Para tanto, utilizou-se o comando identificação de registros duplicados (*Identify Duplicate*) contido no programa *Statistical Package for the Social Sciences* (SPSS) versão 17.0 para *Windows* com as variáveis supracitadas, solicitando a ordenação pela a última data de procedimento/medicamento lançada na APAC/SIA-SUS.

Na base do RCBP-SP, foram excluídos 18 casos duplicados, pois continham o mesmo nome, data de nascimento, topografia e morfologia e haviam realizado a divisão em 24 sub-bases de dados, de acordo com a Classificação Internacional de Doenças Oncológicas (CID-O) 3ª edição (Quadro 1).

A segunda etapa consistiu na padronização das variáveis, utilizadas no *linkage* probabilístico: "nome do paciente", "data de nascimento" e "sexo". O processo de padronização foi realizado para homogeneizar as variáveis das bases de dados, visando minimizar erros no processo de pareamento, como a alocação de registros do mesmo paciente num mesmo bloco lógico, inviabilizando a comparação dos registros, acarretando na classificação de falsos não pares; frequências de escores iguais a zero; erros fonéticos, onde o pareamento dos registros é feito pela comparação aproximada de cadeias de caracteres; e a perda de pares verdadeiros. A estratégia consistiu em remover múltiplos espaços em branco, retirar caracteres diferentes de A-Z (em variáveis específicas), transformar todos os caracteres numéricos em alfanuméricos maiúsculos, suprimir letras duplicadas, remover preposições, omitir os agnomes, transformar o primeiro e último nome em código *soundex*, e substituir letras, por exemplo, W em V, Y em I[6].

Quadro 1. Descrição das 24 sub-bases utilizadas no processo de *linkage*. Registros de Câncer de Base Populacional de São Paulo, 1997 a 2005

Código CID-O 3ª edição	Descrição	Tamanho inicial
C00.0 a C09.9	Partes do lábio, base da língua, outras partes e partes não especificadas da língua, gengiva, assoalho da boca, palato, outras partes e partes não especificada da bica, glândula parótida, outras glândulas salivares maiores e salivares maiores não especificadas e amígdala	9.321
C10.0 a C11.9	Orofaringe e nasofaringe	1.613
C12.0 a C14.8	Seio piriforme, hipofaringe e outras localizações e localizações mal definidas, do lábio, cavidade oral e faringe	1.381
C15.0 a C15.9	Esôfago	5.579
C16.0 a C17.9	Estômago e Intestino Delgado	18.511
C18.0 a C18.9	Cólon	15.684
C19.0 a C21.8	Junção retossigmóide, reto, ânus e canal anal	10.615
C22.0 a C24.9	Fígado e via biliares intra-hepáticas, outras partes e partes não especificadas das vias biliares	4.021
C25.0 a C26.9	Pâncreas e outros órgãos digestivos e localizações mal definidas do aparelho digestivo	5.896
C30.0 a C33.9	Cavidade nasal e ouvido médio, seio da face, laringe e traquéia	6.260
C34.0 a C34.9	Brônquios e pulmões	17.455
C37.0 a C39.9	Timo, coração, mediastino e pleura, outras localizações e localizações mal definidas do aparelho respiratório e dos órgãos intratorácicos	854
C40.0 a C42.4	Ossos, articulações e cartilagens articulares dos membros e de outras localizações e de localizações não especificadas, sistema hematopoiético e reticuloendotelial	11.358
C44.0 a C44.9 (8720 a 8780)	Pele melanoma	5.034
C44.0 a C44.9	Pele não melanoma	70.662
C47.0 a C49.9	Nervos periféricos e sistema nervoso, retroperitônio e peritônio, tecido conjuntivo, subcutâneo e outros tecidos moles	2.830
C50.0 a C50.9	Mama	41.307
C51.0 a C58.9	Vulva, vagina, colo do útero, corpo do útero, útero, ovário, outros órgãos genitais femininos e os não especificados, e placenta	28.359
C60.0 a C60.9/ C62.0 a C63.9	Pênis, testículo e outros órgãos genitais masculinos e os não especificados	2.153
C61.9	Próstata	29.455
C64.0 a C68.9	Rim, pelve renal, ureter, bexiga, outros órgãos urinários	12.343
C69.0 a C72.9	Olho e anexos, meninges, encéfalo, medula espinhal, nervos cranianos e outras partes do sistema nervoso central	7.199
C73.0 a C75.9	Glândula tireóide e supra-renal, outras glândulas endócrinas e estruturas relacionadas	11.696
C76.0 a C80.9	Outras localizações e localizações mal definidas, linfonodos e localização primária desconhecida	23.720

EXECUÇÃO DO *LINKAGE* PROBABILÍSTICO

Para o *linkage* probabilístico, foram realizados os processos de blocagem e pareamento. O processo de blocagem consistiu na criação de blocos lógicos, mutuamente exclusivos, utilizando as chaves de blocagem, primeiro e último nome do paciente transformadas em código fonético (respectivamente, PBLOCO — *soundex* do primeiro nome — e UBLOCO — *soundex* do último nome), mais as variáveis sexo e ano de nascimento. Quanto ao pareamento, os campos escolhidos foram: "nome do paciente", "data de nascimento" e "ano de nascimento". A combinação entre os campos de blocagem e de pareamento resultaram em catorze estratégias de *linkage*.

A partir disto, calcularam-se os escores, resultantes da soma dos campos empregados no processo de pareamento, nos quais os pesos foram construídos baseados nos conceitos de sensibilidade e especificidade de testes diagnósticos (Quadro 2). Para definir os valores de sensibilidade e especificidade na geração dos escores entre a base de dados do RCBP-SP *versus* PRO-AIM; e para a base do RCBP-SP *versus* APAC-SIA/SUS foram utilizadas probabilidades iniciais (sementes) extraídas do estudo de Camargo Jr. e Coeli[20]. Para o cálculo destas probabilidades, as bases de dados do PRO-AIM e APAC-SIA/SUS foram rodadas integralmente, e da base de dados RCBP-SP completa foi retirada uma amostra de 10% de casos, proporcional aos anos analisados e às topografias. Após definidas as probabilidades, foram gerados dois fatores de ponderação (concordância e discordância), calculados como logaritmo de base dois da razão de verossimilhança entre estas.

DEFINIÇÃO DA ESTRATÉGIA DE RELACIONAMENTO

Para definir a melhor estratégia de relacionamento, iniciou-se o processo de *linkage* entre dois arquivos. O primeiro relacionamento foi feito entre as 24 sub-bases de dados do RCBP-SP e a base de dados do PRO-AIM, e o segundo relacionamento entre as 24 sub-bases de dados do RCBP-SP, sem os óbitos identificados no processamento com o PRO-AIM, e a base de dados da APAC-SIA/SUS.

Com o arquivo de relacionamento gerado, foram eliminados os escores de discordância total, isto é, os valores negativos. Ao final, foram criados arquivos para a identificação de pares verdadeiros ou falsos. Para esta classificação, foram conferidas as variáveis de relacionamento "nome do paciente", "data de nascimento" e as variáveis de confirmação "endereço", "causa básica de óbito" e "tratamento/procedimento realizado". Nesta fase, foi realizada a leitura manual de todos os registros pareados, a partir do escore mínimo de concordância, e, a cada par verdadeiro, foi atribuída a nota um (1) e, a cada par identificado como falso, atribui-se a nota zero (0), possibilitando confrontar com o escore calculado e obter os valores de sensibilidade e especificidade do método. Para considerar um par verdadeiro, os registros pareados deveriam, ao menos, apresentar uma das condições descritas no Quadro 3. Ao todo, foram lidas, por três pesquisadores, 648 bases de dados com tamanho máximo de 1.309.283 registros pareados

Após esta fase, definiu-se a blocagem PBLOCO, UBLOCO e SEXO com as variáveis de pareamento "nome" e "data de nascimento" do paciente como a estratégia de relacionamento utilizada para identificação do ponto de corte para pares verdadeiros. Nesta definição, foi considerado o maior número de pares verdadeiros, o menor tempo de execução para o relacionamento, menor número de duplicidades e a precisão do *linkage*, dado pelo escore máximo alcançado.

Foram calculados os valores da área sob a curva (*Receiver Operating Characteristic* – ROC), a sensibilidade e a especificidade para determinar o ponto de corte do escore de maior precisão na identificação dos pares verdadeiros. Estas análises foram realizadas no programa SPSS versão 17.0 para *Windows*.

Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo (protocolo 0086.0.207.000-08) e pela Secretaria Municipal da Saúde (protocolo 0064.0.162.000-09).

Quadro 2. Parâmetros para a construção dos escores de pareamento e fatores de ponderação

Parâmetros e fatores de ponderação	RCBP-SP *versus* PRO-AIM			RCBP-SP *versus* APAC-SIA/SUS		
	Nome	Data de Nascimento	Ano de Nascimento	Nome	Data de Nascimento	Ano de Nascimento
Algoritmo	Aproximado	Caractere	Diferença	Aproximado	Caractere	Diferença
Sensibilidade (m_i)	96,2%	92,5%	95,0%	87,1%	83,0%	89,5%
1-especificidade (u_i)	0,001%	0,001%	6,400%	0,001%	1,000%	5,3%
Proporção mínima de concordância	85,0%	65,0%	± 2	85,0%	65,0%	± 2
Pesos da concordância*	16,89	6,39	3,88	16,62	6,42	4,08
Pesos da discordância	-4,70	-3,72	-4,24	-2,97	-2,54	-3,17

*Valor total do escore RCBP-SP *vs* PRO-AIM "Nome" + "Data de Nascimento" = 23,3 e "Nome" + "Ano de Nascimento" = 20,8; RCBP-SP *vs* APAC-SAI/SUS "Nome" + "Data de Nascimento" = 23,0 e "Nome" + "Ano de Nascimento" = 20,7. RCBP-SP: Registros de Câncer de Base Populacional de São Paulo; PRO-AIM: Programa de Aprimoramento de Mortalidade; APAC-SIA/SUS: Autorização e Procedimentos de Alta Complexidade do Sistema de Informações Ambulatoriais do Sistema Único de Saúde.

RESULTADOS

Observou-se, nesta pesquisa, que as variáveis utilizadas nas etapas blocagem e pareamento no processo de *linkage* probabilístico apresentaram poucos valores ignorados. A base de dados do RCBP-SP não continha casos ignorados nas variáveis "nome do paciente" e "sexo". No entanto, a "data de nascimento" apresentou 35,9% de informações ausentes. Para a base de dados do PRO-AIM, as variáveis "nome do paciente", "data de nascimento" e "sexo" apresentaram, respectivamente, 0,900, 0,800 e 0,003% de valores ignorados. Já para a base de dados da APAC-SIA/SUS, 100% da variável "sexo" estava preenchida, assim como a variável "nome do paciente". Porém, nesta última, foram identificados e retirados caracteres de pontuação excedentes. Em 5.397 casos havia o caractere apóstrofo ('), 4.413 ponto-e-vírgula (;), 1.736 hífen (-) e 5.878 vírgula (,). Para a variável "data de nascimento", o percentual de informações ignoradas foi de 0,1%.

Quadro 3. Condições para considerar um par como verdadeiro

	Ao menos uma das seguintes condições deveria estar presente:
1.	Nome, sobrenome e data de nascimento iguais, em nomes pouco comuns (isto é, João, José, Maria, Aparecida, Santos, Souza, Silva, etc). Os nomes e sobrenomes poderiam apresentar grafias diferentes.
2.	Nome, sobrenome, dia e mês da data de nascimento iguais e o último dígito do ano de nascimento com variação de dois anos para mais ou para menos, em nomes pouco comuns. Os nomes e sobrenomes poderiam apresentar grafias diferentes.
3.	Nome, sobrenome e ano de nascimento iguais, com dia e mês invertidos, com endereço compatível, em nomes pouco comuns. Os nomes e sobrenomes poderiam apresentar grafias diferentes.
4.	Nome e sobrenome iguais, sem data de nascimento, com endereço compatível e causa básica de óbito ou tratamento por câncer, em nomes pouco comuns. Os nomes e sobrenomes poderiam apresentar grafias diferentes.
5.	Nome e sobrenome iguais, sem data de nascimento, com idade compatível e endereço compatível e data do óbito igual, em nomes pouco comuns. Os nomes e sobrenomes poderiam apresentar grafias diferentes.
6.	Nome e três ou mais sobrenome iguais, sem data de nascimento, com idade compatível, endereço compatível, data do óbito igual ou causa básica do óbito ou tratamento por câncer. Os nomes e sobrenomes poderiam apresentar grafias diferentes.

Tabela 1. Valores de sensibilidade, especificidade e área da curva *Receiver Operating Characteristic* para o *linkage* Registros de Câncer de Base Populacional de São Paulo (1997 a 2005) *versus* Programa de Aprimoramento de Mortalidade no ponto de corte escore 18

Topografia	Sensibilidade (%)	Especificidade (%)	Área da curva ROC	N° de pares vinculados (≥18)	N° de pares verdadeiros (≥18)	N° registros da base de dados total	N° de pares verdadeiros da base de dado total
C00.0 a C09.9	94,0	98,9	97,3	2.398	2.378	11.219	2.447
C10.0 a C11.9	93,2	99,2	98,9	790	780	5.056	813
C12.0 a C14.8	93,8	98,9	98,2	649	604	4.906	644
C15.0 a C15.9	96,7	98,7	98,7	2.716	2.597	11.568	2.687
C16.0 a C17.9	93,8	98,7	98,5	7.080	6.915	19.829	7.370
C18.0 a C18.9	89,0	99,3	97,7	5.581	5.451	16.819	6.126
C19.0 a C21.8	93,9	98,4	97,5	3.251	3.111	11.842	3.312
C22.0 a C24.9	92,8	98,6	98,5	2.228	2.140	8.547	2.307
C25.0 a C26.9	95,5	98,7	98,4	3.766	3658	12.072	3.831
C30.0 a C33.9	95,3	99,0	98,1	2.038	1..964	9.577	2.060
C34.0 a C34.9	92,7	98,6	98,0	8.709	8.535	21.869	9.204
C37.0 a C39.9	94,9	98,7	98,0	430	394	3.168	415
C40.0 a C42.4	93,9	98,6	98,1	4.677	4.577	13.580	4.874
C44.0 a C44.9 (8720 a 8780)	73,7	99,4	94,0	624	592	5.745	803
C44.0 a C44.9	81,7	98,7	94,7	3.581	3.362	21.157	4.113
C47.0 a C49.9	94,8	98,7	98,3	856	797	5.220	841
C50.0 a C50.9	90,0	99,0	97,7	6.344	6.291	13.316	6.988
C51.0 a C58.9	97,5	99,3	99,0	9.382	7.537	270.609	7.728
C60.0 a C60.9/ C62.0 a C63.9	92,8	99,2	98,4	468	359	13.246	387
C61.9	92,0	99,0	98,3	9.169	7.110	221.695	7.731
C64.0 a C68.9	94,2	98,7	97,4	3.336	3.218	12.639	3.417
C69.0 a C72.9	94,2	98,5	98,6	2.925	2.808	10.974	2.890
C73.0 a C75.9	85,7	99,1	96,2	672	635	5.077	741

ROC: *Receiver Operating Characteristic.*

Ao utilizar a estratégia de relacionamento PBLOCO + UBLOCO + SEXO com "nome" e "data de nascimento" do paciente, foram estimados os escores na identificação de um par verdadeiro. Na análise das topografias, verifica-se que o ponto de corte no escore 18 apresentou boa acurácia, com valores de sensibilidade que variaram de 73,7 a 96,7%, e de especificidade de 98,5 a 99,4%. Destaca-se que abaixo deste ponto de corte, poucos pares verdadeiros foram encontrados. Na Tabela 1, para o menor valor de sensibilidade observado (73,7%) (sub-base C44.0 a C44.9 morfologia 8720 a 8780, pele melanoma), dos 803 pares verdadeiros, 592 estavam acima ou igual ao escore 18. Observa-se que, neste ponto de corte e acima, a revisão se restringiu a 624 registros pareados, enquanto que a base de dados total desta localização foi formada por 5.745 registros pareados.

Para a sub-base de dados da topografia C51.0 a C58.9 (vulva, vagina, colo do útero, corpo do útero, útero, ovário, outros órgãos genitais femininos e os não especificados e placenta), a estratégia de blocagem relacionou 270.609 registros com sensibilidade 97,5%, e especificidade de 99,3%. Verifica-se que, acima ou igual ao escore 18, foram identificados 7.537 pares verdadeiros em 9.382 registros pareados. Por outro lado, nos 261.227 registros pareados restantes abaixo deste ponto de corte, foram identificados, apenas, 191 pares verdadeiros (Tabela 1).

Ressalta-se que o mesmo ponto de corte (escore ≥18) observado entre o relacionamento das 24 sub-bases do RCBP-SP *versus* PRO-AIM foi encontrado para o *linkage versus* APAC-SIA/SUS. Na Tabela 2, observa-se que os valores de sensibilidade variaram de 56,7 a 99,0%, e de especificidade entre 93,0 a 98,4%. No *linkage* da sub-base C44.0 a C44.9, morfologia 8720 a 8780 (pele melanoma), onde se encontra o menor valor de sensibilidade (56,7%) e de especificidade (98,2%), dos 150 pares verdadeiros, 85 foram identificados acima ou igual ao escore 18. Os demais 65 pares verdadeiros foram localizados nos 4.810 registros pareados restantes.

DISCUSSÃO

Esta pesquisa teve como objetivo propor um ponto de corte nos escores obtidos pelo *linkage* probabilístico e avaliar a acurácia deste método na identificação de pares verdadeiros.

Tabela 2. Valores de sensibilidade, especificidade e área da curva *Receiver Operating Characteristic* para o *linkage* Registros de Câncer de Base Populacional de São Paulo (1997 a 2005) *versus* Autorização e Procedimentos de Alta Complexidade do Sistema de Informações Ambulatoriais do Sistema Único de Saúde no ponto de corte escore 18.

Topografia	Sensibilidade (%)	Especificidade (%)	Área da curva ROC	N° de pares vinculados (≥18)	N° de pares verdadeiros (≥18)	N° de pacientes da base de dados total	N° de pares verdadeiros da base de dados total
C00.0 a C09.9	96,0	97,6	98,9	1.125	557	23.481	580
C10.0 a C11.9	96,9	97,4	99,3	392	84	11.339	87
C12.0 a C14.8	87,7	97,5	94,6	128	50	3.131	57
C15.0 a C15.9	66,9	98,3	94,5	214	97	7.011	145
C16.0 a C17.9	99,0	97,5	99,3	1.046	389	26.226	393
C18.0 a C18.9	81,2	98,4	97,9	648	371	16.963	457
C19.0 a C21.8	83,2	96,2	92,4	554	297	7.076	357
C22.0 a C24.9	83,6	98,7	99,2	227	51	12.704	61
C25.0 a C26.9	80,5	98,6	92,2	106	33	5.215	41
C30.0 a C33.9	98,9	96,8	99,1	962	356	18.826	360
C34.0 a C34.9	80,8	98,0	92,9	603	449	8.245	556
C37.0 a C39.9	81,0	98,2	96,4	53	17	1.976	21
C40.0 a C42.4	98,4	97,3	99,3	1.220	547	24.723	556
C44.0 a C44.9 (8720 a 8780)	56,7	98,2	89,7	170	85	4.981	150
C44.0 a C44.9	80,1	96,2	93,1	1.457	827	17.652	1.032
C47.0 a C49.9	98,7	98,2	99,6	482	223	14.757	226
C50.0 a C50.9	69,6	80,2	81,3	4.890	2380	16.069	3.419
C51.0 a C58.9	88,1	96,4	98,2	3.269	1616	48.192	1.834
C60.0 a C60.9/ C62.0 a C63.9	98,8	98,4	99,6	691	165	32.242	167
C61.9	86,5	98,5	96,5	3.268	1.357	127.734	1.569
C64.0 a C68.9	98,7	96,6	99,0	1.107	304	23.729	308
C69.0 a C72.9	72,8	96,2	89,3	453	201	6.873	276
C73.0 a C75.9	92,7	97,8	99,1	730	253	21.537	273

ROC: *Receiver Operating Characteristic*.

Verificou-se pelos achados que, para todas as topografias, o ponto de corte no escore 18 apresentou melhor desempenho, com áreas sob a curva ROC variando entre 94,7 a 99,0%. Isto reflete a efetividade do método e sua aplicabilidade, dado que os pares verdadeiros foram identificados com maior precisão num menor número de registros pareados e com baixo risco de perdas, podendo este processo ser incorporado à rotina do RCBP-SP, tanto na recuperação quanto na completitude das informações.

Vale lembrar que a vigilância de uma doença ou o monitoramento de qualquer situação de saúde, administrativa ou financeira depende, além de recursos humanos e financeiros, da organização dos serviços, da cobertura do evento a ser medido e da qualidade da informação. Da mesma maneira que o sucesso de um *linkage* depende dos fatores, cobertura e qualidade do dado. A baixa cobertura de um sistema de informação pode implicar na perda de pares verdadeiros, devido à inexistência de registro na fonte de informação, sugerindo à baixa acurácia da técnica. Quanto à qualidade, sabe-se que o processo de *linkage* probabilístico está relacionado ao poder de discriminação das variáveis utilizadas, isto é, pelos valores válidos e sua uniformidade[7,11].

Quanto à cobertura das bases de dados, o RCBP-SP cobre todo o município. Os dados são coletados de forma passiva e ativa em 301 fontes de notificação. A coleta ativa é feita periodicamente em todos os hospitais, clínicas, laboratórios de anatomia patológica, serviço de verificação de óbitos e Registros de Câncer de Base Hospitalar[21,22]. A base do PRO-AIM, segundo o Ministério da Saúde[22,23], apresenta uma cobertura superior a 90% e, embora a cobertura da APAC-SIA/SUS seja referente a procedimentos e medicamentos de pacientes que utilizaram os serviços do Sistema Único de Saúde (SUS), a SMS-SP relata que os procedimentos de alta complexidade registrados nesse sistema, devido ao seu custo elevado e a limitada cobertura pelos planos privados de saúde, representam 90% do total de procedimentos realizados na região Sudeste. Para o Município de São Paulo, verifica-se que a taxa de cobertura pelos planos de saúde é de 59,8% e na pesquisa realizada pelo Instituto de Estudos de Saúde Suplementar (IESS) foi identificado que 26% das pessoas que possuem assistência médica privada também utilizam o SUS[23,24].

Em relação à qualidade das bases de dados, no ano de 2008 a equipe do RCBP-SP fez uma verificação manual de todos os casos incidentes registrados entre 1997 a 2005 com a finalidade de eliminar as possíveis duplicidades. Atualmente, para o controle de qualidade desta base, o RCBP-SP possui dupla digitação dos casos novos e, antes de serem agregados, estes ficam em um banco denominado temporário para que seja feita a revisão por uma supervisora. Somente após estas etapas, o novo caso é agregado ao banco definitivo[21].

Por outro lado, a ausência da variável data de nascimento foi quase 36% e nas presentes, verificou-se que muitas estavam incompletas ou apresentavam registro como 01 de janeiro de um ano referido. Destaca-se que esta foi uma das limitações do processo, pois um número elevado de valores ignorados dificultou o *linkage* dos bancos de dados, levando a não identificação de alguns pares verdadeiros, em que na dúvida o registro pareado foi classificado como falso (não par), fato agravado pela grande quantidade de homônimos existentes nesta base de dados. Além disso, as variáveis de confirmação do *linkage,* como "nome da mãe" e "endereço de residência", também apresentavam baixa completitude, respectivamente, 0,5 e 23,6%. No estudo de Karmel et al.[25], que objetivou descrever os aspectos empíricos da técnica de *linkage*, foi observado que a presença de valores ignorados diminuiu a probabilidade de identificar um par verdadeiro.

Em relação ao PRO-AIM, não foram identificados maiores problemas com a completitude das variáveis. Os registros que apresentavam, no campo "nome do indivíduo que veio a óbito", as palavras "indigente" ou "suspeito" ou outro caractere incomum totalizaram 0,85% da base de dados. Evidencia-se que o CEIINFO, departamento responsável pelo PRO-AIM, tem como uma de suas macrodiretrizes promover a melhoria dos processos de produção da informação em saúde e como diretriz aprimorar a gestão dos SIS-SMS-SP[26].

Os erros de preenchimento da base de dados da APAC-SIA/SUS foram observados com frequência. Observou-se que, a partir do primeiro registro, os demais para aquele mesmo paciente apresentavam falhas na informação, como nomes abreviados, endereço incompleto, CPF zerado ou em branco ou ainda diferentes números deste documento para um mesmo paciente, "nome da mãe" em branco ou com as palavras "O MESMO", fazendo referência ao registro anterior. Outros estudos também mostraram problemas no preenchimento ou ausência de informação em muitas variáveis com campos obrigatório[27,28]. Desta forma, acredita-se que possa ter ocorrido mais de um registro para o mesmo paciente. No entanto, na fase de leitura manual, após o *linkage* de base de dados, foram deixados os pares verdadeiros com a última data devido ao tamanho da base de dados da APAC-SIA/SUS conter, aproximadamente, 32 milhões de registros. Grande parte da sua limpeza, para transformar os registros de procedimentos em pacientes, foi executado de forma automatizada.

A partir do conhecimento sobre as limitações das bases de dados, foi determinado o ponto de corte mais preciso na identificação de um par verdadeiro. A definição do valor foi baseada em uma estratégia de blocagem e pareamento, pois todos os

registros da base de dados foram lidos "um a um". Todavia, a literatura ressalta que, para um melhor desempenho do método probabilístico, recomenda-se que mais de uma estratégia seja empregada, variando de três a cinco passos, pois aumentam a probabilidade de um par verdadeiro ser capturado[16,17,29]. Entretanto, o número de pares verdadeiros encontrados a partir da segunda estratégia de blocagem representou 4 a 10% do total de pares verdadeiros. Além disso, estas estratégias menos restritas aumentam o número de registros falsos relacionados com maior tempo de execução[17,30,31,]

Quanto aos resultados do escore, para todas as topografias, o valor maior ou igual ao escore dezoito foi o mais acurado, dado que acima deste ponto de corte encontram-se a maioria dos pares verdadeiros, otimizando a revisão manual. Este achado corrobora com o estudo de Tromp et al.[8] no qual foi observado que, a partir do escore 18, estavam localizados 94% dos pares verdadeiros[8]. Destaca-se que, no artigo citado, os autores, além do uso do "nome" e "data de nascimento" para o *linkage*, utilizaram o "CEP" e o "hospital" para o cálculo do escore, o que não foi viável, pois as variáveis referentes ao logradouro apresentaram muitos valores ignorados. Vale ressaltar que a digitação desta informação não era obrigatória para o RCBP-SP no período analisado.

Todavia, o menor valor de sensibilidade observado nesta pesquisa foi atribuído à sub-base de dados pele melanoma (topografias C44.0 a C44.9 M8720 a 8780). Observou-se que, nesta sub-base, muitos casos registrados possuíam nomes que apresentavam grafias complexas e fonemas diferentes dos identificados na população brasileira. Neste contexto, os sobrenomes que apresentavam encontros consonantais complexos, possivelmente eram de origem europeia, tornando difíceis de serem compreendidos e, consequentemente, digitados, aumentando a probabilidade de erro. uma vez que sabemos que a qualidade do registro está diretamente ligada ao poder discriminante da variável, acredita-se que, por apresentar grafias pouco habituais, o erro de preenchimento da ficha de coleta e o erro de digitação foram acima do observado para outros tipos de topografias, acarretando valores menores para a sensibilidade e a especificidade no escore sugerido. Para avaliação desses pares, foi necessária a leitura em voz alta por mais de um pesquisador envolvido no estudo, pois esta possibilitou verificar a similaridade dos fonemas.

O elevado número de registros na base de câncer de pele do tipo melanoma com supostos problemas de grafias e fonemas revela que as características das bases de dados devem ser levadas em consideração. Nesta pesquisa, sabendo-se que o risco para o melanoma é ter ascendência ou origem europeia, a probabilidade de se observar nomes pouco comuns foi maior que nas outras topografias[32,33]. O mesmo foi observado para a topografia de câncer de mama (C50.0 a C50.9) no *linkage* entre a base do RCBP-SP e APAC-SAI/SUS. Tal fato refletiu numa menor acurácia para estas bases de dados no ponto de corte determinado. Nestes casos o uso de diferentes padronizações dos sobrenomes, levando em consideração características fonéticas de línguas estrangeiras, como as de origem anglo-saxônicas e anglo-germânicas, podem melhorar o desempenho do escore.

A otimização da identificação de pares verdadeiros pelo *linkage* probabilístico permite que esta técnica seja empregada em diversos estudos na área de pesquisa em saúde pública e epidemiologia, bem como administrativos, viabilizando as relações temporais entre eventos na história do paciente, ampliando seu uso em diversos setores. Ainda assim, conforme a disponibilidade de tempo, recursos humanos dos serviços e objetivos do *linkage*, como a identicação de casos incidentes, recomenda-se que mais de uma estratégia de blocagem e pareamento menos restritas possam ser utilizadas, fazendo uso do mesmo ponto de corte.

CONCLUSÃO

Conclui-se nesta pesquisa que o ponto de corte foi capaz de auxiliar na escolha de um par verdadeiro, com excelente acurácia, apresentando grandes benefícios para tornar rotineiro o processo de *linkage* em grandes bases de dados epidemiológicas. Os resultados mostraram que, acima ou igual ao escore dezoito, encontravam-se quase a totalidade dos pares verdadeiros. Já abaixo deste escore, para a maioria das bases de dados, menos de 1% dos registros pareados eram pares verdadeiros.

REFERÊNCIAS

1. Howe GR. Use of computerized record Linkage in cohort studies. Epidemiol Rev. 1998;20(1):112-21.
2. Fundação Sistema Estadual de Análise de Dados (Seade). Dados para repensar a Aids no Estado de São Paulo: resultados da parceria entre Programa Estadual DST/Aids e Fundação Seade. São Paulo: Fundação Seade; 2010.
3. Soares C, Silva GA. Uso de registro de assistência farmacêutica do Sistema de Informação Ambulatorial para avaliação longitudinal de utilização e adesão a medicamentos. Cad Saúde Colet. 2013;21(3):245-52.
4. Santos SA, Legay LF, Aguiar FP, Lovisi GM, Abelha L, Oliveira SP. Tentativas e suicídios por intoxicação exógena no Rio de Janeiro, Brasil: análise de informações através do *linkage* probabilístico. Cad Saúde Pública. 2014;30(5):1057-66.

5. McGlynn EA, Damberg CL, Kerr EA, Brook RH. Health information systems: design issues and analytic applications. Santa Monica: Rand Health; 1998.

6. Camargo Jr. KR, Coeli CM. *Reclink:* aplicativo para o relacionamento de bases de dados, implementando o método *probabilistic record linkage*. Cad Saúde Pública. 2000;16(2):439-47.

7. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. Cad Saúde Pública. 2004;20(2):362-71.

8. 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(5):565-72.

9. Oberaigner W, Stuhlinger W. Record linkage in the Cancer Registry of Tyrol, Austria. Methods Inf Med. 2005;44(5):626-30.

10. Nitsch D, Morton S, DeStavola BL, Clark H, Leon DA. How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of the 1950s study. BMC Med Res Methodol. 2006;6:15.

11. Méray N, Reitsma JB, Ravelli ACJ, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. J Clin Epidemiol. 2007;60(9):833-91.

12. Fournel I, Schwarzinger M, Binquet C, Benzenine E, Hill C, Quantin C. Contribution of record linkage to vital status determination in cancer patients. Stud Health Technol Inform. 2009;150:91-5.

13. van Herk-Sukel MPP, van de Poll-Franse LV, Lemmens VEPP, Vreugdenhil G, Pruijt JFM, Coebergh JWW, et al. New opportunities for drug outcomes research in cancer patients: the linkage of the Eindhoven cancer registry and the PHARMO record linkage system. Eur J Cancer. 2010;46(2):395-404.

14. Fleming M, Kirby B, Penny KI. Record linkage in Scotland and its applications to health research. J Clin Nurs. 2012;21(19-20):2711-21.

15. 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(10):2249-52.

16. 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(7):1431-38.

17. Migowski A, Chaves RBM, Coeli CM, Ribeiro ALP, Tura BR, Kuschnir MCC, et al. Accuracy of probabilistic record linkage in the assessment of high-complexity cardiology procedures. Rev Saúde Pública. 2011;45(2):269-75.

18. Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. 1969;64(328):1183-210.

19. Jaro MA. Advances in Record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 1989;84:414-20.

20. Camargo Jr KR, Coeli CM. Reclink III: relacionamento probabilístico de registros. Versão 3.1.6.3160. Rio de Janeiro; 2007.

21. Brasil. Ministério da Saúde. Secretaria de Estado da Saúde. Registro de Câncer de Base Populacional de São Paulo. Câncer em São Paulo 1997-2008: incidência, mortalidade e tendência de câncer no Município de São Paulo. São Paulo, S.P. RCBP-SP; 2011.

22. Brasil. Ministério da Saúde. A experiência brasileira em sistemas de informação em saúde. Brasília: Editora do Ministério da Saúde, 2009,49-70.

23. Agência Nacional de Saúde Suplementar; *ANS 2011* Cadernos de Informação de Saúde Suplementar: Beneficiários, Operadoras e Planos. Available from: http://www.ans.gov.br/images/stories/Materiais_para_pesquisa/Perfil_setor/Caderno_informacao_saude_suplementar/2011_mes06_caderno_informacao.pdf

24. Brasil. Ministério da Saúde. Instituto de Estudos de Saúde Suplementar. Os custos do ressarcimento ao SUS. Saúde Suplementar em Foco. Informativo Eletrônico IESS. 2010;1(8).

25. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. BMC Health Serv Res. 2010;10:41.

26. Coordenação de Epidemiologia e Informação (CEInfo). Sistema de monitoramento e avaliação da qualidade das bases de dados - SUS: histórico, conceitos, indicadores e métodos. São Paulo; 2011.

27. Brito C, Portela MC, Vasconcellos MTL. Avaliação da concordância de dados clínicos e demográficos entre autorizações de procedimentos de alta complexidade oncológica e prontuários de mulheres atendidas pelo sistema único de saúde no estado do Rio de Janeiro, Brasil. Cad Saúde Pública. 2005;21(6):1829-35.

28. Cherchiglia ML, Guerra Jr. AA, Andrade ELG; Machado CJ, Acúrcio FA, Meira Jr. W, et al. A construção da base de dados nacional em terapia Renal Substitutiva (TRS) centrada no indivíduo: aplicação do método de *linkage* determinístico-probabilístico. Rev Bras Estud Popul. 2007;24(2):163-7.

29. Coeli CM, Camargo Jr. KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol. 2002;5(2):185-96.

30. Nakhaee F, McDonald A, Black D, Law M. A feasible method for linkage studies avoiding clerical review: linkage of the national HIV/AIDS surveillance databases with the national death index in Australia. Aus N Z J Public Health. 2007;31(4):308-12.

31. Taniguchi MT, Pelaquin MHH, Latorre MRDO. Relacionamento probabilístico entre as bases de dados do registro de câncer de São Paulo e do sistema de informações de mortalidade municipal. [Trabalho de conclusão de curso]. São Paulo: Universidade de São Paulo; 2006.

32. Bradford PT. Skin cancer in skin of color. Dermatol Nurs. 2009;21(4):170-7.

33. Gonçalves FT, Francisco G, de Souza SP, Luiz OC, Festa-Neto C, Sanches JA, et al. European ancestry and polymorphisms in DNA repair genes modify the risk of melanoma: a case-control study in a high UV index region in Brazil. J Dermatol Sci. 2011;64(1):59-66.

Recebido em: 07/10/2014
Aprovado em: 05/01/2015