Daniele Pinto da Silveira[I]

Elizabeth Artmann[II]

Acurácia em métodos de relacionamento probabilístico de bases de dados em saúde: revisão sistemática

Accuracy of probabilistic record linkage applied to health databases: systematic review

RESUMO

OBJETIVO: Analisar a literatura nacional e internacional sobre validade de métodos de relacionamentos nominais de base de dados em saúde, com ênfase nas medidas de aferição da qualidade dos resultados.

MÉTODOS: Revisão sistemática de estudos de coorte, caso-controles e seccionais que avaliaram a qualidade dos métodos de relacionamento probabilístico de base de dados em saúde. Foi utilizada metodologia Cochrane para revisões sistemáticas. As bases consultadas foram as mais amplamente utilizadas: Medline, LILACS, Scopus, SciELO e Scirus. Não foi utilizado filtro temporal e os idiomas considerados foram: português, espanhol, francês e inglês.

RESULTADOS: As medidas sumárias da qualidade dos relacionamentos probabilísticos foram a sensibilidade, a especificidade e o valor preditivo positivo. Dos 202 estudos identificados, após critérios de inclusão, foram analisados 33 artigos. Apenas seis apresentaram dados completos sobre as medidas-sumárias de interesse. Observam-se como principais limitações a ausência de revisor na avaliação dos títulos e dos resumos dos artigos e o não-mascaramento da autoria dos artigos no processo de revisão. Estados Unidos, Reino Unido e Nova Zelândia concentraram as publicações científicas neste campo. Em geral, a acurácia dos métodos de relacionamento probabilístico de bases de dados variou de 74% a 98% de sensibilidade e 99% a 100% de especificidade.

CONCLUSÕES: A aplicação do relacionamento probabilístico a bases de dados em saúde tem primado pela alta sensibilidade e uma maior flexibilização da sensibilidade do método, mostrando preocupação com a precisão dos dados a serem obtidos. O valor preditivo positivo nos estudos aponta alta proporção de pares de registros verdadeiramente positivos. A avaliação da qualidade dos métodos empregados tem se mostrado indispensável para validar os resultados obtidos nestes tipos de estudos, podendo ainda contribuir para a qualificação das grandes bases de dados em saúde disponíveis no País.

DESCRITORES: Sistemas de Informação. Modelos Estatísticos. Gerenciamento de Informação. Bases de Dados Estatísticos. Relações Interinstitucionais. Revisão.

[I] Programa de Pós-Graduação em Saúde Pública. Escola Nacional de Saúde Pública Sergio Arouca (ENSP). Fundação Oswaldo Cruz (Fiocruz). Rio de Janeiro, RJ, Brasil

[II] Departamento de Administração e Planejamento. ENSP. Fiocruz. Rio de Janeiro, RJ, Brasil

Correspondência | Correspondence:
Elizabeth Artmann
Escola Nacional de Saúde Pública Sérgio Arouca
R. Leopoldo Bulhões, 1480, 7o andar – Manguinhos
21041-210 Rio de Janeiro, RJ, Brasil
E-mail: artmann@ensp.fiocruz.br

Recebido: 12/10/2008
Revisado: 8/4/2009
Aprovado: 15/4/2009

ABSTRACT

OBJECTIVE: To analyze both national and international literature on validity of record linkage procedure of health databases focusing on quality assessment of results.

METHODS: A systematic review of cohort, case-control, and cross-sectional studies that evaluated quality of probabilistic record linkage of health databases was conducted. Cochrane methodology of systematic reviews was used. The following databases were widely searched: Medline, LILACS, Scopus, SciELO and Scirus. A time filter was not applied and articles were searched in the following languages: Portuguese, Spanish, French and English.

RESULTS: Summary measures of the quality of probabilistic record linkage were sensitivity, specificity, and positive predictive value. There were identified 202 studies, and after applying the inclusion criteria, a total of 33 articles were reviewed. Only six had complete data on the summary measures of interest. The main limitations were: no reviewer to evaluate titles and abstracts; and no blinding of the article's authors in the review process. Most scientific publications in this field were from the United States, United Kingdom, and New Zealand. Overall, the accuracy of probabilistic record linkage of databases ranged from 74% to 98% sensitivity and 99% to 100% specificity.

CONCLUSIONS: Probabilistic record linkage of health databases has notably been characterized by high sensitivity and greater flexibility of the procedure's sensitivity, indicating concern with data accuracy. The positive predictive value in studies shows a high proportion of truly positive record pairs. The quality assessment of these procedures has been proved essential for validating the results obtained in these studies, and can also contribute to improve large health databases available in Brazil.

DESCRIPTORS: Information Systems. Models, Statistical. Information Management. Statistical Databases. Interinstitutional Relations. Review.

INTRODUÇÃO

O número de estudos voltados ao desenvolvimento e aprimoramento de métodos de relacionamento nominal de bases de dados vem crescendo desde os anos 1980, sendo a maior parte dos trabalhos conduzidos e publicados nos EUA, Reino Unido e Nova Zelândia.[1,5,6,12] No Brasil, apesar de uma extensa difusão e aplicação deste método em estudos de diversas áreas de conhecimento, em especial na epidemiologia, ainda são poucos os trabalhos que visam a identificar um mesmo indivíduo em duas ou mais bases de dados nominais.

Nas últimas décadas, importantes sistemas nacionais de informação foram desenvolvidos pelo Ministério da Saúde do Brasil, com notáveis avanços na disseminação eletrônica de dados sobre nascimentos, óbitos, doenças de notificação, atendimentos hospitalares e ambulatoriais, atenção básica e orçamentos públicos em saúde, entre outros.[a] No entanto, a diversidade de concepção dos sistemas e a ausência de um identificador unívoco que integre as várias bases tornam o trabalho de relacionamento nominal de bases de dados no Brasil bastante complexo. A produção e a utilização de informações em saúde no País se processam em um complexo contexto de relações institucionais, compreendendo variados mecanismos de gestão e financiamento. Portanto, a integração das diversas bases de dados também ocorre de forma descontínua e desarticulada entre os vários níveis e atores governamentais interessados nesse tema.

O relacionamento de bases de dados (*record linkage*) pode ser definido como uma área do conhecimento voltada ao estudo do método de busca de pares ou registros duplicados dentro de um mesmo arquivo ou entre arquivos. O relacionamento nominal de bases de dados é usualmente realizado por meio de métodos probabilísticos que empregam processos de pareamento de duas (ou mais) bases utilizando probabilidades de concordância e discordância entre um conjunto de variáveis

[a] Organização Panamericana de Saúde. Rede Interagencial de Informações para Saúde (Ripsa). Indicadores básicos para a saúde no Brasil: conceitos e aplicações. Brasília; 2002.

comuns às duas bases.[1] Normalmente, são usadas para a identificação de indivíduos variáveis como: nome, endereço e data de nascimento. Informações adicionais como renda, educação, entre outras, podem ser utilizadas, dependendo da qualidade destes campos.[a]

O principal objetivo do relacionamento probabilístico de base de dados é encontrar pares de registros que se referem a uma mesma pessoa, bem como padronizar e verificar a qualidade das informações.[10,b] Todavia, a qualidade dos dados constantes nos sistemas de informação em saúde pode dificultar o processo de relacionamento das bases ou contribuir para erros no pareamento das variáveis. Deste modo, é fundamental avaliar a precisão dos métodos empregados para relacionamento de registros médicos nominais, estatísticas vitais e grandes bases de dados nacionais para aprimorar a capacidade de identificar registros individuais e aumentar a qualidade e fidedignidade das informações. Uma das formas de avaliar a precisão dos métodos de relacionamento nominal de bases de dados é por meio de estudos de acurácia.

Na epidemiologia, a acurácia é considerada uma medida de validade muito aplicada em estudos sobre avaliação de testes diagnósticos ou de rastreamento.[7,14] Os estudos de acurácia permitem avaliar em que grau os dados medem o que eles deveriam medir ou o quanto os resultados de uma aferição correspondem ao verdadeiro estado do fenômeno aferido. A acurácia de um teste ou de um método é comumente considerada em relação a algum padrão-ouro. Entretanto, nem sempre é possível obter um padrão-ouro para que tal comparação seja estabelecida, seja em termos de um novo teste clínico, seja em termos de um novo método a ser empregado. Em alguns casos é preciso escolher como padrão de validade outro método que é reconhecidamente imperfeito.[7] Sendo assim, para avaliar a performance dos métodos de relacionamento nominal de bases de dados, também é necessário comparar os resultados obtidos no processo de relacionamento com uma fonte de informação independente sobre a ocorrência dos desfechos de interesse (padrão-ouro). Contudo, a disponibilidade dessas fontes costuma ser restrita.[8] Quando não há padrão-ouro para determinar a especificidade e a sensibilidade do relacionamento das bases de dados, a qualidade do *linkage* pode ser avaliada apenas por meio de medidas indiretas. Alguns autores reportam o uso dessas medidas[1,16] como, por exemplo, no estudo conduzido por Blakely et al (1999)[c] em que a proporção de registros do banco de mortalidade relacionados com registros da outra base em cada etapa do processo de relacionamento foi utilizada para estimar o número de falsos-positivos usando o número de links duplicados nas duas bases.

A sensibilidade e a especificidade são as medidas tradicionais de validade empregadas quando a exposição e o desfecho são variáveis categóricas. Em epidemiologia, sensibilidade refere-se à proporção de pessoas que apresentam o desfecho de interesse que são classificadas como positivas no teste enquanto a especificidade refere-se à proporção de pessoas que não possuem a doença ou desfecho de interesse e são identificadas como negativas pelo teste. Nos estudos de precisão de relacionamento nominal de base de dados, a acurácia pode também ser aferida em termos de sensibilidade, especificidade e valor preditivo positivo do método, dado que, por analogia, o par-verdadeiro do *linkage* (*match*) pode ser considerado equivalente à presença do desfecho de interesse nos estudos epidemiológicos (e.g. óbito).[1,14]

O presente estudo teve por objetivo analisar a literatura nacional e internacional sobre validade de métodos de relacionamentos nominais de base de dados em saúde, com ênfase nas medidas utilizadas para aferir a qualidade dos resultados.

PROCEDIMENTOS METODOLÓGICOS

Procedeu-se a uma revisão sistemática retrospectiva de estudos de coorte, caso-controles e seccionais que tiveram como principal objetivo mensurar a acurácia (precisão) dos métodos empregados no relacionamento probabilístico de base de dados em saúde. Foi utilizada metodologia de revisões sistemáticas proposta pela Colaboração Cochrane.[9] A identificação dos artigos foi feita por busca bibliográfica nas bases de dados Medline, LILACS, Scirus, SciELO e Scopus, entre novembro e dezembro de 2007. A não utilização de filtro por ano de publicação se deveu ao fato de supormos que haveria poucas publicações e, portanto, poderíamos perder artigos importantes. O único limitador foi o próprio filtro do Medline que apenas busca artigos a partir de 1966.

As estratégias e palavras-chave selecionadas para busca nas bases de dados foram: (*record linkage*) OR (*record linkage* AND *health*) OR (*record linkage* AND *accuracy*) OR (*record linkage* AND *health information*) OR (*record linkage* AND *information system*) OR (*record linkage* AND *accuracy* AND *probabilistic* AND *specificity* AND *health data*) OR (*record linkage* AND

[a] Winkler WE. Automatically Estimating Record Linkage False Match Rates. Washington: Statistical Research Division United States Census Bureau; 2007.

[b] Shah GH, Taima F, McBrida S. Probabilistic linkage in Public Health: Results of the NAHDO Survey. A critical assessment of record linkage software used in public health. Salt Lake City: National Association of Health Data Organizations; 2008[citado 2008 jan 21]. Disponível em: http://www.nahdo.org/.

[c] Blakely T, Salmond C, Woodward A. Anonymous record linkage of 1991 census records and 1991-94 mortality records: The New Zealand Censu-Mortality Study. Wellington: Department of Public Health, School of Medicine, University of Otago; 1999.

health AND *accuracy*) OR (*medical record linkage*) OR (*medical record linkage* AND *accuracy*) OR (*medical record linkage* AND *health*). Foram selecionados artigos publicados nos idiomas português, espanhol, inglês e francês.

Posteriormente, uma segunda estratégia utilizada foi busca manual em listas de referência dos artigos identificados e selecionados.

Os estudos tiveram como desfecho as medidas de acurácia do método de relacionamento probabilístico: sensibilidade, especificidade e valor preditivo positivo. Assim, como nos estudos epidemiológicos que usualmente se valem dessas medidas para verificar a acurácia de testes clínicos, diversos autores têm aplicado tais conceitos ao campo do relacionamento de registros.[1,5,6,12,16] A sensibilidade nos estudos de acurácia foi definida como a proporção de todos os registros em um arquivo ou base de dados que tenha um *match* em outro arquivo corretamente aceito como *link* (verdadeiros-positivos). Já a especificidade refere-se à proporção de todos os registros em um arquivo que *não* tenha *matchs* em outra base corretamente não aceitos como *links* (verdadeiros-negativos). O valor preditivo positivo foi mensurado em estudos de acurácia por meio da ocorrência de *links* duplicados (exemplo: um registro de óbito relacionado a dois ou mais registros de outra base de comparação). Essa ocorrência pode ser usada para estimar o valor preditivo positivo do resultado dos critérios de classificação (modelo de decisão).

Na literatura internacional, o *match* é a definição utilizada para referir-se a um par de registros que pertence ao mesmo indivíduo, enquanto o *link* é um par que é aceito como provável par "verdadeiro" na etapa da blocagem dos registros, quando são criados pequenos blocos de registros que serão comparados entre si.[1]

Não houve um revisor na avaliação dos títulos e dos resumos de todos os estudos localizados na busca eletrônica e a autoria dos artigos não foi ocultada no processo de revisão.

Quanto aos critérios de inclusão e exclusão, inicialmente, foi verificado se cada estudo apresentava a maioria dos principais critérios de inclusão: tipo de desenho, universo do estudo (ou tamanho da amostra), tamanho das bases, sensibilidade, especificidade e valor preditivo positivo. Utilizou-se um formulário padronizado para coletar as informações dos artigos.

Foram excluídos da análise os artigos referentes a métodos determinísticos de relacionamento de bases de dados.

Foi utilizada uma amostra por conveniência seguindo os critérios de inclusão e exclusão definidos para o estudo.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A busca bibliográfica resultou em 180 artigos, destes 47 foram selecionados com base no título e na leitura do resumo. Apenas 33 artigos foram selecionados para revisão, pois 14 foram excluídos por não apresentarem os critérios previamente definidos no estudo. Quatro estudos analisados foram identificados nas referências bibliográficas dos artigos. Todos os trabalhos foram categorizados segundo tipo de estudo. Os estudos seccionais representaram 54% (n=18) de todos os artigos revisados, enquanto os de coorte, 24% (oito) do total. A maioria dos estudos de coorte identificada foi classificada como coorte retrospectiva (n=7).

A maioria dos estudos apresentou uma alta sensibilidade do método, em torno de 93% a 99%, embora estudos com sensibilidades menores também tenham sido encontrados (74% a 83%).

Após a leitura na íntegra dos artigos, observou-se que poucos estudos apresentavam dados sobre a especificidade (n=5) e o valor preditivo positivo (n=4) do método de *linkage* empregado. A maioria dos artigos (n=28) apenas apresentava informações sobre os valores de sensibilidade do método, pois a aferição das demais medidas poderia ser de difícil mensuração. Observou-se que, na maioria dos estudos, isso apenas era possível em três situações: a) disponibilidade de outra base de dados correlata que pudesse validar os dados identificados por meio do método probabilístico, b) quando o pesquisador tinha disponíveis resultados de estudos epidemiológicos (como, por exemplo, coortes ou estudos transversais) que auxiliem na comparação dos resultados e na estimativa do número de falsos-negativos (pares que eram verdadeiros e foram classificados como falsos) e c) por meio da estimativa de parâmetros que permitissem inferir o valor da especificidade do método (aplicação de métodos de estatística bayesiana). Deste modo, devido a essas limitações, optou-se por apresentar em detalhe apenas os artigos que continham informação completa sobre os valores obtidos para sensibilidade, especificidade e valor preditivo positivo (Tabela).

Os artigos foram agrupados por período no intuito de favorecer uma análise do volume de trabalhos ao longo dos anos. A partir do ano 2000 observou-se maior volume de publicações. Quanto ao universo, a maioria dos estudos (n=31) utilizou bases populacionais ou hospitalares, além de registros administrativos em saúde (n=6) (Figura).

De acordo com os artigos selecionados para revisão, os países que mais publicaram sobre a avaliação de métodos de relacionamento probabilístico de base de dados foram: Estados Unidos (n=8), Nova Zelândia (n=6) e Reino Unido (n=6). Houve quatro trabalhos do Brasil. Os demais trabalhos foram publicados em países da Europa.

Tabela. Características dos estudos analisados sobre acurácia da metodologia do relacionamento probabilístico de base de dados em saúde de 1999 a 2006.

Autor/Ano	País	Tipo de estudo	Tamanho das bases	Sensibilidade	Especificidade	Valor preditivo positivo
Ellekjaer et al, 1999	Noruega e Suécia	seccional	70.000 registros de base populacional de doença cerebrovascular e 759 registros de altas hospitalares	86%	não descrita	68%
Blakely & Salmond, 2002	Nova Zelândia	seccional	3.131.176 registros do New Zealand *Census-Mortality Study* e 39.515 registros de mortalidade de 1986 a 1989	não descrita	não descrita	93% a 99%
Grannis et al, 2003	EUA	seccional	2 pares de arquivos com 6.000 registros do *Social Security Death Master File*	99,2% (1° Registro)	99,4% (1° Registro)	não descrito
				99,0% (2° Registro)	99,4% (2° Registro)	
Zingmond et al, 2004	Reino Unido	coorte retrospectivo	1.858.458 registros de alta hospitalar da Califórnia	95%	99%	99%
			69.757 registros de óbitos hospitalares (1990 a 1999)			
Coutinho & Coeli, 2006	Brasil	coorte prospectivo	250 registros hospitalares de uma coorte de idosos internados por fratura no município do Rio de Janeiro e registro de óbitos do estado do RJ (n do registro de óbitos não informado)	86%	99%	98%
Nagle et al, 2006	Nova Zelândia	coorte retrospectivo	822 registros de mulheres com diagnóstico de câncer de ovário entre 1990 e 1993 do Index Nacional de Óbitos (NDI/Australia), registros de câncer de base populacional e 450 óbitos do NDI	93%	100%	não descrito

Quanto ao tamanho das bases, foi possível observar que a maioria dos trabalhos utilizou bases de porte médio, variando de 100.000 a 700.000 registros. Muitos trabalhos referem-se também ao uso de pequenas bases de dados de estatísticas vitais com um tamanho amostral de três a quatro dígitos.

A pesquisa com uma das maiores bases observadas foi a de Victor & Mera,[15] ao relacionar 1,7 milhões registros administrativos de seguros-saúde dos Estados Unidos a 8,5 milhões de registros sobre assistência à saúde, obtendo uma sensibilidade de 92% no processo de *linkage*.

O programa mais amplamente utilizado no Brasil para a realização destes estudos foi o RecLink.[3] Nos outros países, observa-se uma variação no uso de distintos programas de relacionamento de base de dados, como o Automach®; o *Statistics Canada's Generalized Record Linkage System*, desenvolvido pelo governo canadense, além de programação computacional.

Dentre os estudos que avaliaram a especificidade do método, os valores variaram entre 99% e 100%. Já os valores preditivos positivos situaram-se entre 68% e 99%.

Os estudos conduzidos por Blakely & Salmond[1] e Zigmond et al[17] apresentaram os maiores valores

preditivos positivos. Zigmond et al relacionaram por meio de programação computacional 1.858.458 registros de alta hospitalar com 69.757 registros de óbitos hospitalares, tendo também obtido alta sensibilidade (95%) e alta especificidade (99%) do método. No Brasil, em pesquisa realizada por Coutinho & Coeli,[5] utilizando bases de dados bem menores (n=250 registros hospitalares), um resultado semelhante foi encontrado com especificidade e valor preditivo positivo elevados, 99% e 98,10%, respectivamente.

CONSIDERAÇÕES FINAIS

O relacionamento probabilístico de base de dados tem sido amplamente utilizado na saúde pública nos últimos 50 anos, desde a publicação do trabalho de Newcombe et al.[11] Ao longo das últimas décadas, o reconhecimento da importância de avaliar os resultados obtidos por meio dos métodos de *linkage* nominal tem crescido e diversos estudos foram conduzidos na tentativa de estimar a precisão destes métodos.

O processo e os métodos de relacionamento de base de dados nem sempre são suficientes para prover as informações necessárias para o pesquisador decidir quando um par de registros é de fato um par verdadeiro (*match*). Nessas situações, necessitam ser utilizadas informações adicionais, além das providas pelas variáveis de pareamento. A revisão manual dos registros é a escolha mais comum dos investigadores, considerada na literatura internacional o padrão-ouro nesta modalidade de estudo. Contudo, nem sempre esta é uma opção viável devido ao tamanho das bases de dados usualmente utilizadas na pesquisa em saúde, tornando em algumas situações o processo de validação dos resultados do *linkage* um processo trabalhoso e caro.[10]

No caso dos estudos de acurácia dos métodos de relacionamento nominal, uma dificuldade é nem sempre ser possível encontrar uma base de dados de referência que possa ser utilizada como parâmetro de comparação e confirmação do *status* de um dado registro em outra base. Além desta dificuldade inicial, a qualidade das bases de dados interfere no processo de relacionamento e é um elemento adicional que deve ser considerado no planejamento de estudos desta natureza, uma vez que os campos a serem utilizados no processo de pareamento das variáveis em muitos casos têm problemas de preenchimento ou de validade.

Na maioria dos estudos, a acurácia do processo de relacionamento probabilístico é fortemente dependente do número e da qualidade dos campos disponíveis para a comparação. Poucos campos disponíveis podem aumentar a ocorrência de pares falso-positivos, pares classificados como verdadeiros que, entretanto, referem-se a pessoas diferentes nas bases de dados comparadas.[6]

De um modo geral, no presente estudo, observou-se que o tamanho das bases não estava necessariamente ligado ao resultado obtido na sensibilidade do método. Esse achado parece confirmar o pressuposto de que a acurácia do método está mais diretamente relacionada à qualidade dos registros e dos campos que serão utilizados para o relacionamento probabilístico.

Além disso, autores como Brenner et al[2] apontam que os erros homônimos tendem a aumentar com o aumento do número de registros nas bases utilizadas para o *linkage*.

Camargo & Coeli[3] observam que a diminuição do valor preditivo positivo nos processos de relacionamento de bases de maior tamanho pode estar associada a fatores como a redução da prevalência dos pares verdadeiros.

Alguns autores apontam que a medida mais adequada para mensurar a qualidade de um processo de relacionamento de bases é o valor preditivo positivo.[2,4] Porém, como é praticamente impossível rever todos os pares de registros, a solução seria gerar uma amostra

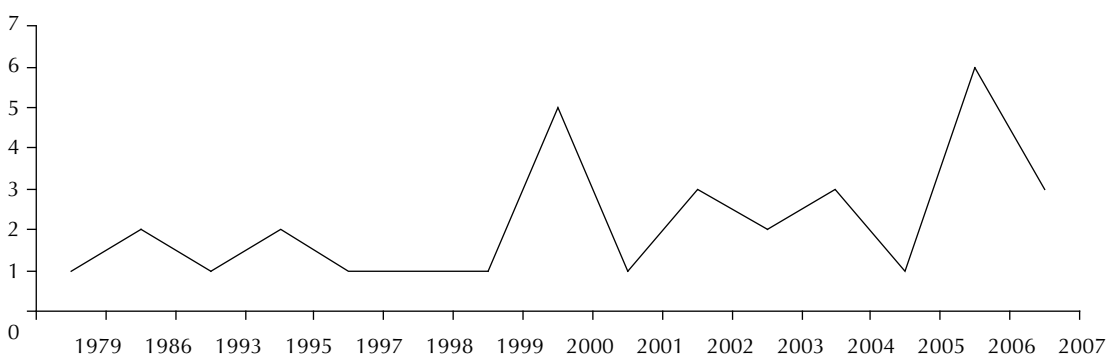

Figura. Série temporal das publicações científicas sobre acurácia do relacionamento probabilístico em bases de dados de 1979 a 2007.

e rever apenas alguns pares. Sauleau et al[13] observam que, apesar de esta ser uma solução viável, não atende completamente ao objetivo de avaliar a qualidade do processo de geração dos dados, resultante do relacionamento das bases.

O indicador mais adequado para apurar a acurácia do relacionamento probabilístico de bases de dados seria então o percentual de registros duplicados.[13] Blakely & Salmond[1] também já haviam apontado que a ocorrência de *links* duplicados no processo de *linkage* pode ser utilizada para quantificar o valor preditivo positivo do método.

Para outros autores, a sensibilidade do método de *linkage* também pode ser estimada adotando-se como padrão-ouro o total de pares verdadeiros identificados tanto durante a busca automática, por meio de programas específicos, como durante a busca manual. Ainda, em alguns casos, intervalos de confiança podem ser adicionados na estimação.[3]

No estudo desenvolvido por Camargo & Coeli,[3] as sensibilidades do processo manual e do processo automático foram muito semelhantes nas situações em que foram utilizadas bases com menor número de registros. Todavia, à medida que as bases aumentavam de tamanho, foi observada uma diminuição do processo de revisão manual, mas não do processo de relacionamento automático das bases.

Portanto, quando não há nenhum padrão-ouro para determinar a especificidade e a sensibilidade do relacionamento das bases de dados, a qualidade do *linkage* pode ser avaliada apenas por meio de medidas indiretas. Algumas destas medidas foram utilizadas no estudo conduzido por Blakely et al[a] como, por exemplo, o percentual de registros de uma base de dados identificados em outra base em cada etapa do processo de relacionamento e percentual de registros da base de dados identificados em outra base ao final do processo de relacionamento.

Para os autores,[a] esse percentual total se aproxima da sensibilidade do relacionamento probabilístico, que seria igual ao número de *links* verdadeiros (*matchs*) identificados dividido pelo número total de pares verdadeiros. A especificidade seria o número de *links* incorretos rejeitados, dividido pelo número total de *links* incorretos.

O método descrito por Blakely & Salmond[1] para estimar o número de falsos-positivos (*Duplicate Method*) é aplicável somente quando há um par verdadeiro (*match*) para um dado registro – situação comum em estudos epidemiológicos, como por exemplo, relacionamento de bases sobre mortalidade com outras bases. Este método quantifica o número de falsos-positivos acima de um dado peso total utilizando o número de *links* duplicados observados acima desse *score* de pontuação. A ocorrência de *links* duplicados pode ser utilizada para quantificar o valor preditivo positivo. Essa quantificação permite uma decisão informada sobre o peso do ponto de corte acima do qual os *links* serão aceitos.

Para Grannis et al,[8] o método probabilístico de base de dados representa um avanço sobre o método determinístico por uma série de razões. Uma delas é a elevação da sensibilidade em torno de 7% com um pequeno decréscimo na especificidade. A sensibilidade de métodos determinísticos, embora possa atingir 100%, pode diminuir significativamente quando são utilizados dados com diferentes características de identificação, como diferentes nomes étnicos, dentre outros.

Em face de todas as informações apresentadas, reitera-se a relevância da utilização e do aprimoramento dos métodos de relacionamento probabilístico de bases de dados no campo da saúde coletiva. A avaliação da qualidade dos métodos empregados tem se mostrado indispensável para validar os resultados obtidos nestes tipos de estudos, podendo ainda contribuir para a qualificação das grandes bases de dados em saúde disponíveis no País. Mais estudos sobre a acurácia dos métodos de *linkage* em pesquisas epidemiológicas são necessários no Brasil.

[a] Blakely T, Salmond C, Woodward A. Anonymous record linkage of 1991 census records and 1991-94 mortality records: The New Zealand Censu-Mortality Study. Wellington: Department of Public Health, School of Medicine, University of Otago; 1999.

REFERÊNCIAS

1. Blakely T, Salmond C. Probabilistic Record Linkage and a method to calculate the positive predictive value. *Int J Epidemiol*. 2002,31(6):1246-52. DOI: 10.1093/ije/31.6.1246
2. Brenner H, Schmidtmann I, Stegmaier C. Effects of record linkage errors on registry-based follow-up studies. *Stat Med*. 1997;16(23):2633-43. DOI: 10.1002/(SICI)1097-0258(19971215)16:23<2633::AID-SIM702>3.0.CO;2-1
3. Camargo Jr KR, Coeli CM. RecLink: aplicativo para o relacionamento de base de dados, implementando o método probabilistic record linkage. *Cad Saude Publica*. 2000;16(2):439-47. DOI: 10.1590/S0102-311X2000000200014
4. Coeli CM, Blais R, Costa MCE, Almeida LM. Probabilistic linkage in household survey on hospital care usage. *Rev Saude Publica*. 2003;37(1):91-9. DOI: 10.1590/S0034-89102003000100014
5. Coutinho ESF, Coeli CM. Acurácia da metodologia de relacionamento probabilístico de registros para identificação de óbitos em estudos de sobrevida. *Cad Saude Publica*. 2006;22(10):2249-52. DOI: 10.1590/S0102-311X2006001000031
6. Coutinho RGM, Coeli CM, Faerstein E, Chor D. Sensibilidade do linkage probabilístico na identificação de nascimentos informados: Estudo Pró-Saúde. *Rev Saude Publica*. 2008;42(6):1097-100. DOI: 10.1590/S0034-89102008005000053
7. Fletcher R, Fletcher S. Epidemiologia clínica: Elementos essenciais. 4.ed. Porto Alegre: Artes Médicas; 2004.
8. 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.
9. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions 4.2.6. Chichester: John Wiley & Sons; 2006[citado em 2007 ago 04]. (The Cochrane Library, 4). Disponível em: http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.pdf
10. Machado CJ, Hill K. Probabilistic Record Linkage and an automated procedure to minimize the undecided-matched pair problem. *Cad Saude Publica*. 2004;20(4):915-25. DOI: 10.1590/S0102-311X2004000400005
11. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. *Science*. 1959;130:954-9. DOI: 10.1126/science.130.3381.954
12. Roos LL, Wajda A. Record linkage strategies. Part I: estimating information and evaluation approaches. *Methods Inf Med*. 1991;30(2):117-23.
13. Sauleau EA, Paumier JP, Buemi A. Medical Record Linkage in health information systems by approximate string matching and clustering. *BMC Med Inform Decis Mak*. 2005;5:32. DOI: 10.1186/1472-6947-5-32
14. Sklo M, Nieto FJ. Epidemiology: Beyond the Basics. London: Jones and Bartlett Publishers; 2004.
15. Victor TW, Mera RM. Record linkage of health care insurance claims. *J Am Med Inform Assoc*. 2001;8(3):281-8.
16. Winkler WE. Record linkage: overview of recent developments and applications. In: Falorsi P, Pallara A, Russo A, editors. L'integrazione di dati di fonti diverse, Technice e applicaczioni del Record Linkage e metodi di stima basati sull'uso congiunto di fonti statistche e amministrative. Rome: Franco Angeli Editore; 2005.
17. Zingmond DS, Ye Z, Ettner SL, Liu H. Linking hospital discharge and death records accuracy and sources of bias. *J Clin Epidemiol*. 2004;57(1):21-9. DOI: 10.1016/S0895-4356(03)00250-6

Trabalho apresentado na *V Annual Meeting HTAI 2008 - Health Technology Assessment in Context*, Montreal, Canadá, 2008.