Aplicação da metodologia de relacionamento probabilístico de base de dados para a identificação de óbitos em estudos epidemiológicos

The use of a probabilistic record linkage methodology in databases to identify death records in epidemiological studies

Resumo

A crescente disponibilidade de dados de abrangência nacional, organizados em distintos sistemas de informação, requer o desenvolvimento de metodologias para o relacionamento de variáveis constantes em diferentes bases de dados. Este artigo descreve e analisa a metodologia utilizada no relacionamento das bases de dados nacionais do Sistema de Informação de Beneficiários (planos privados de assistência à saúde) e do Sistema de Informação de Mortalidade. Foram analisados os registros de óbitos e os registros de beneficiários no ano de 2004, identificando-se 92.566 óbitos em beneficiários de planos privados de saúde. O rigor na especificidade, em detrimento da sensibilidade do método empregado, não gerou vieses de seleção que pudessem comprometer as análises resultantes. A razão de mortalidade padronizada aponta a subestimação do número de óbitos, além de evidenciar diferenças no risco de morte entre as populações analisadas; no entanto, as diferentes situações de vida e saúde a que estão expostas podem ter interferido nos resultados.

Palavras-chave: Relacionamento de bases de dados. Sistemas de informação em saúde. Planos e seguros privados de saúde. Mortalidade.

Juliana Pires Machado
Daniele Pinto da Silveira
Isabela Soares Santos
Márcia Franke Piovesan
Ceres Albuquerque
Agência Nacional de Saúde Suplementar/ Ministério da Saúde

Correspondência: Juliana Pires Machado, Avenida Augusto Severo, 84 - 10º andar, Glória, Rio de Janeiro, RJ, E-mail: juliana.machado@ans.gov.br

O projeto obteve aprovação do Comitê de Ética em Pesquisa da ENSP,nº de processo 16/08,CAAE - 0024.0.031.000-08.

Abstract

The increasing availability of nationwide data organized in distinct information systems requires the development of linkage methodologies to identify the relationships between the same variables in different databases. This article describes and analyzes the linkage methodology between two countrywide databases: the Brazilian Mortality System and the database of the population covered by private health insurance. Death registrations and the records of individuals covered by private health insurance in 2004 were analyzed. 92,566 deaths were identified in those covered by private health insurance. The strictness in specificity instead of sensitivity in the method employed did not lead to selection biases that could have compromised the resulting analyses. The standardized mortality ratio (SMR) indicates underestimation of the number of deaths, in addition to showing differences in the risk of death between the populations analyzed; however, different life and health situations may have affected the results.

Keywords: Record linkage. Health information systems. Private health insurance. Mortality.

Introdução

A crescente disponibilidade de dados de abrangência nacional e a informatização das bases que monitoram eventos de saúde proporcionam vasto campo de estudos, muitas vezes dependentes de dados oriundos de mais de um sistema de informação. A análise destes dados, organizados em distintos sistemas, tem exigido nos últimos anos o desenvolvimento de métodos de *linkage*, ou seja, de ligação ou relacionamento de variáveis de diferentes bases de dados, visando sua interoperabilidade[1-3].

Pesquisas em saúde pública de todo o mundo vêm empregando o relacionamento de bases de dados em diversos tipos de estudos[4,5], a exemplo dos etiológicos[6-9] os de coorte[10,11], os de caso-controle[12,13], e as avaliações de serviços de saúde[14].

O relacionamento de bases de dados é produto da metodologia criada para encontrar um registro pertencente à mesma entidade, constante em dois ou mais sistemas de informação. Dependendo da concordância entre as variáveis das distintas bases, o método utilizado pode ser determinístico ou probabilístico. Quando os registros de cada base possuem variáveis comuns para as quais é possível obter concordância exata, o método de relacionamento é denominado de Relacionamento Determinístico. Sem a presença de um identificador unívoco nas bases a serem relacionadas, a opção é utilizar vários campos/variáveis comuns às duas/ou mais bases e trabalhar com as probabilidades de concordância e discordância entre as variáveis selecionadas para o pareamento. Nesse caso, o método é denominado Relacionamento Probabilístico[15].

No Brasil, há uma ampla utilização do método probabilístico de identificação, pois, especialmente na área da Saúde, ainda não há um número ou código de identificação unívoca do indivíduo, a exemplo do proposto no Cartão Nacional de Saúde (CNS) pelo Ministério da Saúde (MS). Nesse contexto, o método probabilístico traz

como vantagens a agilização e a melhoria da acurácia do processo de relacionamento[16].

A necessidade de se utilizar dados originados de diversos sistemas de informação para construir análises que considerem a complexidade dos contextos humanos impulsiona pesquisas sobre as diversas metodologias de relacionamento probabilístico de bases de dados. Os estudos nesta área incluem desde os estritamente metodológicos até os originados das aplicações do método na investigação de determinantes do adoecimento e vigilância de eventos em saúde.

Instituições e pesquisadores nacionais e internacionais voltam-se para o desenvolvimento e o aperfeiçoamento do método de *linkage*[1,20-25]. Parte importante destes trabalhos tem por objetivo aperfeiçoar o modelo matemático de *linkage* desenvolvido por Fellegi e Sunter em 1969[21], enquanto outros estudos buscam desenvolver inovações na plataforma de algoritmos de *linkage*, que utilizam, na sua maioria, as variáveis nome, sexo e data de nascimento, a exemplo do estudo de Li *et al*, de 2006[5]. Menos comum é a utilização de apenas campos textuais para pareamento, como o realizado por Ravikumar e Cohen em 1969[26], ressaltando-se que a aplicação desta técnica não é adequada para situações em que não há um preenchimento qualificado dos campos das bases em análise.

A maioria dos estudos descreve como etapas do método de relacionamento probabilístico de base de dados: I) padronização dos registros para um mesmo dicionário de preenchimento; II) blocagem para agilização do processo a partir da criação de blocos lógicos; III) pareamento dos registros para definição dos candidatos a pares; e IV) submissão a uma regra de decisão que indica o par de registros com maior probabilidade de ser o verdadeiro[13,19,27-29].

Ainda assim, nem sempre é possível assumir um par encontrado como "verdadeiro" ou "falso"[28]. A solução, neste caso, é a introdução de uma terceira categoria, a de "possível par", correspondendo aos "duvidosos", o que torna o processo mais complexo à medida que aumenta o número de registros a ser relacionado. Aplicando a regra proposta por Jaro[22], na qual atribui-se valor proporcional ao grau de concordância dos campos, resta ao pesquisador definir qual o nível aceitável de discordância, que ditará o ponto de corte para a classificação dos pares na aplicação da regra de decisão.

Outros estudos de extrema relevância para o tema apresentam metodologias de avaliação da técnica empregada para relacionar as bases de dados. São abordagens quanto ao cálculo do valor preditivo positivo em relacionamentos probabilísticos[18] e à avaliação da acurácia através de estudos de seguimento de coortes[25]. Uma das recomendações discutidas a partir dos resultados desses estudos é a utilização de regras de classificação que minimizem a ocorrência dos falsos-positivos, principalmente quando se tratar de grandes bases de dados onde os erros atribuíveis a homônimos tendem a aumentar.

Uma parcela importante da produção científica nesse campo refere-se também a estudos nos quais o relacionamento probabilístico de bases de dados é utilizado na pesquisa epidemiológica[10,13,16,28,30-32]. A maioria desses estudos caracteriza-se por serem ecológicos, casos-controle ou coortes sobre determinantes de mortalidade, estimativas de casos de doenças transmissíveis e avaliações do risco de morbidade e mortalidade por doenças crônicas.

Outra linha de investigação dos estudos sobre relacionamento de bases de dados refere-se à aplicação do método para a vigilância epidemiológica, principalmente com o objetivo de estimar casos subnotificados[33-36].

Este estudo teve como objetivo analisar e descrever o método de relacionamento de bases de dados empregado para identificar óbitos entre a população coberta por planos privados de saúde no Brasil. Os resultados obtidos a partir da aplicação do método foram analisados e estão publica-

dos no capítulo 5 do livro "Saúde Brasil 2006: uma análise da situação de saúde no Brasil"[37].

Metodologia de Relacionamento do Sistema de Informação de Mortalidade (SIM) e do Sistema de Informação de Beneficiários (SIB)

Fontes de dados

SIB: A partir deste sistema, selecionou-se a base de beneficiários de planos de saúde no Brasil com contrato ativo em algum período do ano de 2004, cobertos por plano de saúde médico-hospitalar e, portanto, aptos a receber atendimento na rede credenciada. Em março de 2006, competência de atualização utilizada neste estudo, a base continha 32.414.290 registros de vínculos ativos em 2004, com 37 variáveis cada um.

Gerido pela Agência Nacional de Saúde Suplementar (ANS), este sistema é atualizado mensalmente com dados encaminhados pelas operadoras de planos privados de saúde registradas, que podem informar inclusões, exclusões ou alterações de registros de seus beneficiários. Trata-se de um cadastro de vínculos, podendo um mesmo indivíduo ser beneficiário de mais de um plano e, portanto, constar no sistema tantas vezes quantos forem seus vínculos a planos privados de saúde.

A qualificação do SIB vem ocorrendo desde sua implantação, com a introdução de um processo de submissão dos dados a críticas de consistência e comparações com outros sistemas de informação nacionais para validação do conteúdo, além da orientação periódica para as operadoras sobre o preenchimento dos campos.

Comparações com a população coberta por plano privado de saúde no Brasil segundo a Pesquisa Nacional por Amostra de Domicílios[38], indicam uma superestimação de aproximadamente 7% no SIB do mesmo período, conseqüência da duplicidade de vínculos.

SIM: Fonte de dados sobre os registros de óbitos ocorridos no país, no ano de 2004 contava com 1.021.356 registros, cada um contendo 77 variáveis. O acesso à base se deu a partir do acordo de parceria de trabalho firmado nos ofícios de nº 11493/2003/DIDES/ANS/MS e 12050/2003/DIDES/ANS/MS.

Este sistema foi criado pelo Ministério da Saúde em 1975 para a consolidação regular de dados nacionais sobre mortalidade. É gerido pela Secretaria de Vigilância em Saúde (SVS/MS) e alimentado pelas secretarias municipais e estaduais de saúde com base na Declaração de Óbito (DO). Está embasado legalmente na Lei 6.015, de 31 de dezembro de 1973, alterada pela Lei nº 6.216, de 30 de junho de 1975[39].

A qualidade dos dados registrados no SIM tem melhorado nos últimos anos, e sua cobertura tem sido bem próxima de 100% nas regiões Centro-Oeste, Sudeste e Sul do país[40].

A utilização dos dados dos dois sistemas de informação seguiu as recomendações da Comissão Nacional de Ética em Pesquisa (CONEP), quanto à preservação do sigilo.

Aplicação do Método

Preparação das Bases

Visando a redução de erros na fase de pareamento, foi realizada uma preparação prévia das bases quanto à exclusão de registros e quanto à padronização e codificação de seus campos. Todo este processamento foi realizado utilizando-se a linguagem SQL sobre base Oracle.

Foram *excluídos* registros do SIM duplicados, considerando-se nesta situação o registro com mesmo número de óbito na mesma UF ou com mesmo nome, nome da mãe, data de nascimento e UF. Para a melhoria da qualidade de campos de "nome" utilizados no processo (nome e nome da mãe), foram excluídas informações ou caracteres acrescidos indevidamente ao nome da pessoa e substituídos os padrões identificados de abreviação na

digitação do nome; foram também excluídos os registros cujo campo não correspondia a um nome próprio, incluindo-se aqueles onde tal campo era ignorado.

Por se tratar de um banco de dados com número elevado de registros, a duplicidade de indivíduos no SIB foi tratada no modelo de decisão, que, por se aplicar sobre um menor número de registros, permite maior agilidade no processamento.

Para a padronização dos campos foram aplicadas as mesmas codificações de sexo e formatos de data, além da fonetização dos nomes, visando tornar associáveis os campos dos diferentes bancos de dados. Foi utilizado o padrão de fonetização de nomes da Caixa Econômica Federal (disponível para download em <http://www.datasus.gov.br/ccssis/tfonetiz.htm>), que originalmente era aplicado pelo Datasus no relacionamento das bases de dados do SIB/ANS e do Sistema de Informações Hospitalares (SIH/SUS), com fins de ressarcimento ao SUS dos procedimentos realizados no sistema público por beneficiários que tinham cobertura para tais serviços por plano de saúde.

Blocagem e Pareamento

Com o objetivo de reduzir o número de comparações entre registros e otimizar o processo de pareamento, foram criados blocos lógicos de registros através de programação em Delphi sobre base Oracle.

Nesta fase foram comparados os campos "sexo", "data de nascimento" e os sete primeiros caracteres do "nome fonetizado" (concatenado em primeiro nome + último nome + nome do meio). Caso estes três atributos fossem compatíveis nas duas bases, era selecionado o candidato a par.

Classificados os candidatos a par por blocos lógicos, atribuiu-se pontuação que alcançou no máximo 125 pontos, definida segundo o nível de concordância dos campos: nome fonético completo, nome da mãe fonetizado e UF de residência. Pares com nota final menor que 80 foram descartados, enquanto pares com 80 ou mais pontos seguiram para a próxima etapa (Figura 1).

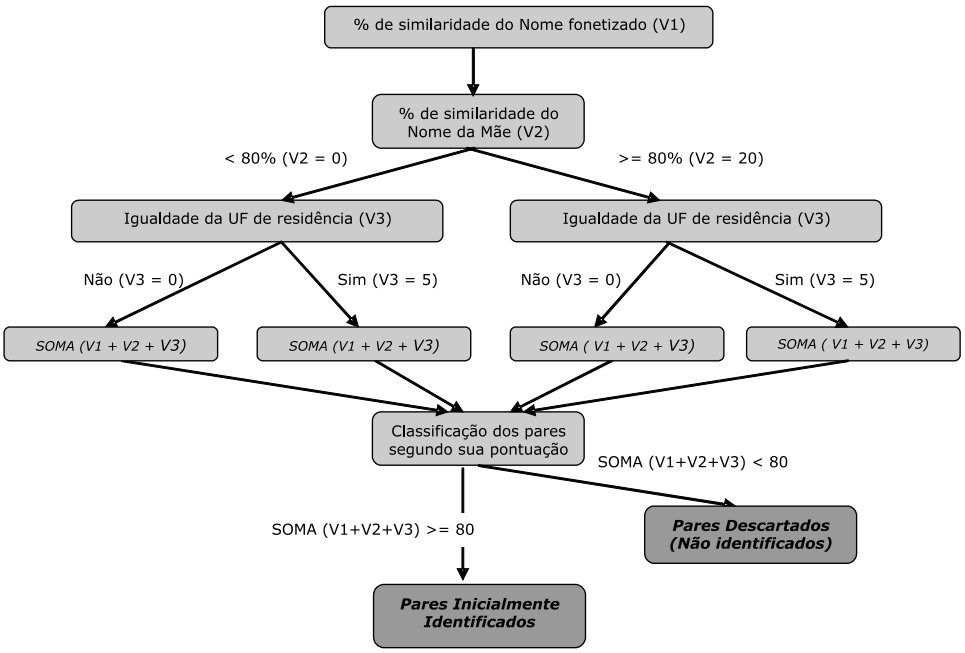

Figura 1 - Esquema de pontuação dos pares inicialmente identificados
***Figure 1** – Scoring system of the pairs initially identified*

Modelo de Decisão

Todos os pares de registros inicialmente identificados na fase anterior e com pontuação a partir de 80 são, nesta etapa, submetidos a uma regra para seleção daqueles com maior chance de serem verdadeiros, utilizando-se a linguagem SQL sobre base Oracle. Três situações geradas na fase de pareamento são possíveis: (1) 1 óbito para 1 beneficiário - 1x1; (2) 1 óbito para N beneficiários - 1xN; (3) N óbitos para 1 beneficiário - Nx1. A situação (2) pode decorrer da possibilidade de duplicidade de vínculos no SIB. Para a situação (1) o par é automaticamente eleito; para cada uma das outras duas situações é aplicada uma seqüência de regras, até que se tenha como produto um par de 1x1. A Figura 2 ilustra a regra de decisão definida para este estudo.

Análise dos Resultados

Para avaliar o método aplicado, tomou-se como medida indireta da precisão a razão de mortalidade padronizada (RMP), utilizando-se como número observado de óbitos o identificado a partir do relacionamento probabilístico das bases e número esperado de óbitos em beneficiários aquele obtido a partir da padronização pelo método indireto, tendo como referência as taxas específicas por idade e sexo no Brasil.

Resultados

Padronização dos Dados

Dos 1.021.356 registros de óbito em 2004, aproximadamente 2% foram excluí-

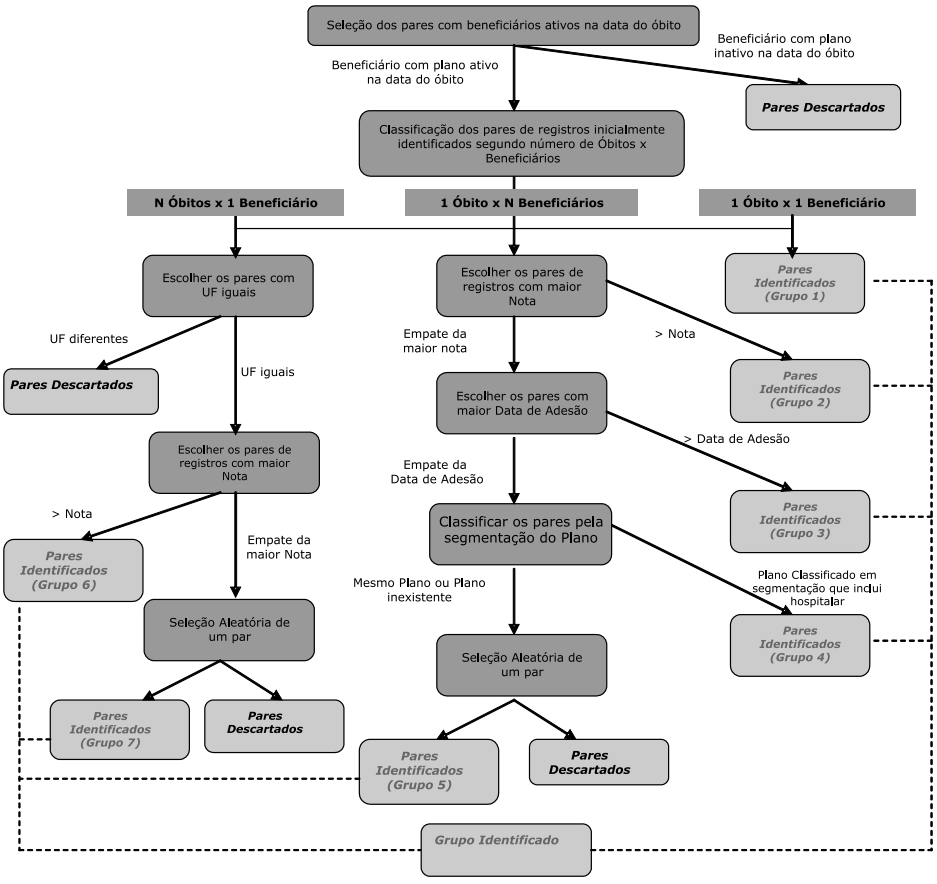

Figura 2 - Etapas da Regra de Decisão entre os pares inicialmente identificados
Figure 2 *- Stages of the Decision Rule between pairs initially identified*

dos do processo de relacionamento das bases de dados após a fase de padronização dos atributos. Destes, a grande maioria foi excluída por erros de preenchimento que impossibilitaram a comparação e apenas 0,05% eram registros duplicados.

Dentre os motivos de exclusão de registros por erro de preenchimento, aquele referente à inclusão da expressão "recém-nascido" ou similar no campo nome foi responsável por mais de 60% do total. Nomes registrados em caracteres não alfabéticos somam quase 30% dos excluídos, enquanto aqueles registrados com a palavra "desconhecido" chegam a aproximadamente 10%. Campos de nome com registro da palavra "natimorto" formam pouco mais de 1% dos excluídos.

Quanto às alterações aplicadas nos campos, o atributo "nome da mãe" deteve o maior número de modificações, sendo até quase 4% corrigido, enquanto o campo "nome" não alcançou 1% de alterações. A data de nascimento não apresentou grande número de alterações e os campos referentes a sexo e UF passaram por modificações em 100% dos registros, uma vez que para o primeiro há padronização do código em um único, e para o segundo há conversão do código do município para se encontrar o referente ao estado (Tabela 1).

Modelo de Decisão

O grupo final somou 92.566 registros. Destes, a grande maioria (89%) foi selecionada na situação referente aos pares 1 óbito x 1 beneficiário (grupo 1). Na situação 1 óbito x N beneficiários, foram identificados aproximadamente 10% dos óbitos de beneficiários. Poucos indivíduos pertencentes à terceira situação, N óbitos x 1 beneficiário, foram identificados (menos de 1%) e destes, a maior parte foi originada de desempate por nota (Tabela 2).

Mortalidade em Beneficiários

Do total de óbitos registrados no SIM no ano de 2004, esperava-se que aproximadamente 23% (ou 234 mil) pertencessem a beneficiários de planos privados de saúde, caso esta população estivesse exposta às mesmas taxas de mortalidade da população brasileira.

O relacionamento probabilístico das bases identificou 92.566 óbitos na população coberta por planos, o que representa 11% do total de óbitos no Brasil, apesar da cobertura por planos privados de assistência médica alcançar 23% da população do país em 2004.

Dos óbitos esperados para a população de beneficiários de planos privados de saúde, caso possuíssem as mesmas taxas da população do Brasil, 39% foram observados (RMP=0,39). O detalhamento destes resultados por sexo, faixa etária e grandes regiões do Brasil pode ser observado nas Tabelas 3 e 4.

Observou-se que a menor RMP encontra-se na faixa etária de menores de 1 ano, em ambos os sexos, o que provavelmente se relaciona à não obrigatoriedade de cadastro

Tabela 1 - Status de preenchimento e alteração dos campos padronizados
***Table 1** – Levels of completeness and changes in standardized fields*

Campo	Situação					
	Preenchido		Nulo		Alterado	
	(n)	*(%)*	*(n)*	*(%)*	*(n)*	*(%)*
Nome da Mãe	986.896	96,626	34.460	3,374	36.531	3,702
Nome	1.021.356	100,000	0	0,000	5.257	0,515
Data de Nascimento	1.011.025	98,990	10.331	1,011	20	0,002
Sexo*	1.021.356	100,000	0	0,000	1.021.356	100,000
UF**	0	0,000	0	0,000	1.021.356	100,000

* Alterado para código único / *Modified to single code.*
** Informação gerada a partir do código do município / *Information generated from the city code.*

Tabela 2 - Seleção segundo Grupos do Modelo de Decisão
Table 2 *- Selection according to Decision Model Groups*

Situação após o pareamento	Grupo	Seqüência de regras do Modelo de decisão	(n)	%
Situação 1: 1 óbito para 1 beneficiário - 1x1	1	Identificação direta	82.917	89,58
Situação 2: 1 óbito para N beneficiários - 1xN;	2	Maior nota	3.050	3,29
	3	Empate anterior + Maior data de adesão	4.193	4,53
	4	Empate anterior + Segmentação que inclua cobertura hospitalar	448	0,48
	5	Empate anterior + Seleção aleatória	1.889	2,04
Situação 3: N óbitos para 1 beneficiário - Nx1	6	UFs iguais e Maior nota	53	0,06
	7	Empate anterior + Seleção aleatória	16	0,02
Todos Situações 1, 2 e 3		Todas as regras	92.566	100,00

Tabela 3 - Razão de Mortalidade Padronizada em Beneficiários de planos de saúde, segundo sexo e faixa etária – Brasil, 2004.
Table 3 - *Standardized Mortality Ratio in individuals covered by private health insurance, by gender and age – Brazil, 2004*

	Total Beneficiários	Masculino	Feminino
< 1 ano	0,16	0,16	0,17
1 a 4 anos	0,40	0,38	0,42
5 a 9 anos	0,43	0,40	0,49
10 a 14 anos	0,47	0,47	0,46
15 a 19 anos	0,40	0,36	0,51
20 a 24 anos	0,32	0,30	0,40
25 a 29 anos	0,30	0,29	0,32
30 a 39 anos	0,31	0,30	0,34
40 a 49 anos	0,33	0,32	0,35
50 a 59 anos	0,37	0,37	0,37
60 a 69 anos	0,43	0,46	0,40
70 a 79 anos	0,44	0,50	0,38
80 anos e mais	0,42	0,48	0,39
Total	0,39	0,41	0,38

de indivíduos até um mês de idade no SIB. Nas outras faixas de idade, observou-se menor RMP entre os indivíduos em faixas economicamente ativas, tendo-se encontrado no sexo feminino as razões mais elevadas entre óbitos identificados e esperados, quando comparado ao sexo masculino.

A mortalidade proporcional por causas em beneficiários é bem similar àquela observada no Brasil. As doenças do aparelho circulatório e as neoplasias possuem altas participações percentuais, tanto na população brasileira quanto na população de beneficiários, observando-se entre os últimos que tais causas são ainda mais importantes. As causas externas destacam-se com menor participação percentual em beneficiários se comparados à população brasileira, assim como as causas mal-definidas, que no Brasil são proporcionalmente quase três vezes mais que em beneficiários (Figura 3).

Uma discussão detalhada acerca do perfil de mortalidade em beneficiários de planos privados de saúde e no Brasil está disponível na publicação "Saúde Brasil 2006: Uma análise da desigualdade em saúde"[37].

Discussão

Apesar da exclusão de registros do SIM devido a erros de preenchimento ou duplicidade da informação, o número de registros utilizados no processo de relacionamento atingiu quase 98% dos dados da base original, com uma perda bastante re-

Tabela 4 - Óbitos registrados no SIM, óbitos esperados e identificados e RMP na população de beneficiários de planos privados de saúde - Brasil, 2004
Table 4 *- Deaths registered in SIM, SMR and expected and identified deaths in the population covered by private health insurance - Brazil, 2004*

Regiões	SIM	Óbitos esperados em beneficiários*	Óbitos identificados em beneficiários	RMP
Centro-oeste	62.545	11.441	3.731	0,33
Nordeste	255.579	29.665	10.583	0,36
Norte	53.630	5.431	2.007	0,37
Sudeste	486.799	162.363	67.008	0,41
Sul	162.803	25.446	9.237	0,36
Brasil	*1.021.356*	*234.369*	*92.566*	*0,39*

* Calculado caso a população de beneficiários se expusesse às mesmas taxas de mortalidade da população brasileira, mantendo sua própria estrutura etária / *Calculated if the population covered by private health insurance had the same mortality ratio as the Brazilian population, keeping its own age structure.*

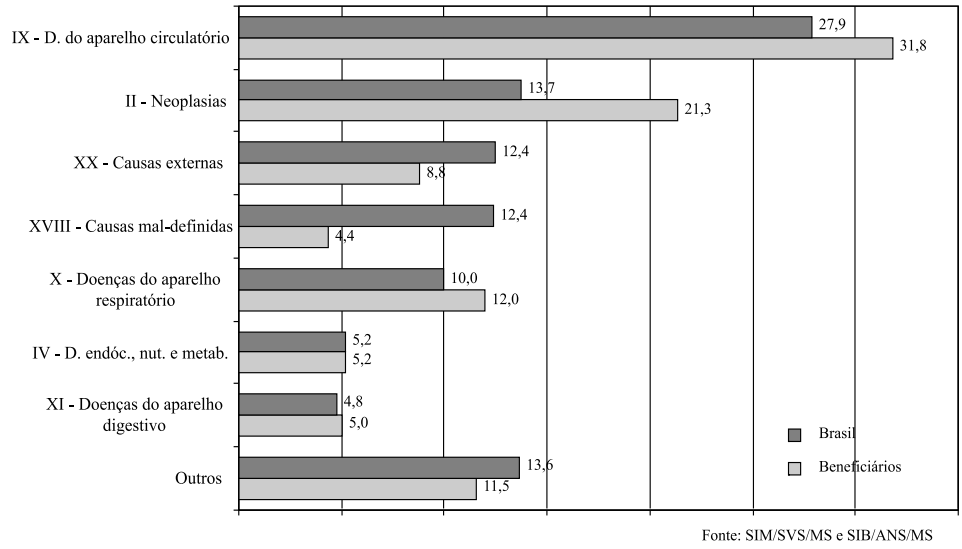

Figura 3 - Mortalidade proporcional por grupos de causa segundo população - Brasil, 2004
Figure 3 *- Mortality ratio by causes of death according to population – Brazil, 2004*

duzida de indivíduos. Dos 2% excluídos do relacionamento, apenas 37% tinham campos de nome preenchidos com a palavra "desconhecido" ou com caracteres alfanuméricos, e seriam candidatos a indivíduos beneficiários de planos de saúde. Os outros 63% teriam pouca oportunidade de participar da saúde suplementar como beneficiários, uma vez que eram natimortos ou recém-nascidos (estes últimos legalmente cobertos pelo plano da mãe até 1 mês de idade, sem obrigatoriedade de registro no SIB, o que interfere diretamente na sua identificação e contribui para a subestimação dos resultados na faixa etária). Com isso, destaca-se que o método utilizado na limpeza e padronização dos campos foi bem sucedido, não interferindo de forma relevante nas perdas observadas.

Se considerada a interpretação teórica para a medida de RMP encontrada a partir dos dados obtidos com o método descrito, poderíamos inferir que a população de beneficiários de planos de saúde (RMP = 0,39) está submetida a um menor risco de morte que a população brasileira. No entanto, deve-se observar que além de dife-

rentes situações de saúde, outro fator que pode contribuir para este resultado é a subestimação do número de óbitos entre os beneficiários, inerente ao processo de relacionamento probabilístico. Não se sabe em que medida o resultado obtido é influenciado por cada um dos seguintes fatores: qualidade das bases utilizadas (SIM e SIB), acurácia do método de relacionamento probabilístico, e a própria condição de vida e saúde das populações do Brasil e de beneficiários de planos de saúde.

Ainda assim, mesmo sem análises quanto à acurácia do método aplicado neste estudo, os resultados obtidos quanto ao perfil de mortalidade da população de beneficiários foram coerentes com o padrão esperado, dada a composição etária (proporcionalmente mais idosos do que a população brasileira, o que se relaciona ao maior percentual de mortes por doenças crônico-degenerativas) e a inserção social (maior renda e inserção no mercado de trabalho, o que se relaciona ao menor risco de morrer por causas externas) deste grupo[37]. Assim, podemos inferir que o erro implícito ao processo probabilístico não apresentou vieses de seleção de indivíduos que pudessem comprometer as análises proporcionais resultantes.

Embora tais resultados mostrem-se consistentes, o cálculo de taxas pode ser prejudicado pois o método de relacionamento probabilístico utilizado privilegia a especificidade em detrimento da sensibilidade, uma vez que foi originalmente criado para identificar indivíduos com o objetivo de cobrança de ressarcimento ao SUS. Em estudo de acurácia do relacionamento probabilístico de bases de dados, Coutinho e Coeli também estimaram alta especificidade e menor sensibilidade, respectivamente, 99 e 85%. Seus resultados comparam-se aos de outros autores, canadenses e escoceses, que encontraram inclusive percentuais de sensibilidade mais elevados[25].

Considerando-se a alta acurácia estimada em outros estudos de relacionamento de bases de dados que utilizam metodologia similar à deste estudo[25], além dos resultados consistentes encontrados no perfil de mortalidade, pode-se inferir que tal diferença entre o esperado e o observado não decorre privativamente do método de relacionamento, mas também da própria diferença de condição de vida entre os indivíduos beneficiários de planos privados de saúde e a população brasileira.

As diferenças no acesso aos serviços de saúde, renda familiar, escolaridade, nível cultural, moradia, trabalho, entre outros fatores, indicam distintas condições sociais e traduzem o impacto do contexto social sobre a saúde. Embora ainda não seja possível mensurar com exatidão o quanto cada fator é determinante da mortalidade, a análise comparada na população de beneficiários de planos privados de saúde e na população brasileira deve, também, considerar a influência desse conjunto de variáveis.

Considerações Finais

O relacionamento probabilístico aplicado para a identificação de óbitos em beneficiários gerou resultado próximo ao esperado quanto ao perfil de mortalidade. Embora a análise dos resultados obtidos permita inferir sobre a precisão do método aplicado, seria necessário a utilização de medidas de acurácia para se conhecer com detalhes os efeitos do erro implícito no processo de relacionamento probabilístico das bases, que também sofre influência da qualidade de preenchimento dos dados. Só então devem ser revistos os critérios de blocagem e pareamento, com o objetivo de aumentar a sensibilidade, que pode contribuir na redução da subestimação dos óbitos encontrados.

Apesar de a mortalidade proporcional ter apresentado resultados próximos ao esperado, o cálculo de medidas de risco não é adequado utilizando-se os dados obtidos neste relacionamento de bases de dados, uma vez que é provável que o numerador esteja subestimado.

Considerando a acurácia descrita por

outros autores que trabalharam com métodos de relacionamento probabilístico, provavelmente ainda resta uma parcela considerável de óbitos esperados, porém não encontrados, que se devem a outros fatores que não o erro implícito ao método, como a qualidade das bases de dados e a própria diversidade da situação de saúde dos beneficiários em relação à população brasileira.

Referências

1. Winkler WE. *Data Quality: Automated Edit/Imputation and Record Linkage*. Washington, DC: Statistical Research Division, U.S. Bureau of the Census; 2006. Disponível em http://www.census.gov/srd/www/byyear.html [Acessado em 12de abril de2007].

2. Bruin A, Kardaun J, Gast F, Bruin E, Van Sijl M, Verweij G. Record linkage of hospital discharge register with population register: Experiences at Statistics Netherlands. *Stat J NU* ECE 2004; 21: 23-32.

3. Winkler WE. *Overview of Record Linkage and Current Research Directions*. Washington, DC: Statistical Research Division, U.S. Bureau of the Census; 2006 Disponível em http://www.census.gov/srd/www/byyear.html [Acessado em 09 de janeiro de 2007].

4. 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 Coletiva* 2006; 14(2): 197-224.

5. Li B, Quan H, Fong A, Lu M. Assessing Record linkage between health care and vital statistics databases using deterministic methods. *BMC* 2006; 6: 48. Disponível em http://www.biomedcentral.com/1472-6963/6/48 [Acessado em 29de janeiro de 2007].

6. 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 1950s study. *BMC* 2006; 6: 15.

7. Whiteman D, Murphy M, Hey K, O'Donnell M, Goldacre MJ. Reproductive factors, subfertility, and risk of neural tube defects: a case-control study based on the Oxford record linkage study register. *Am J Epidemiol* 2000; 152: 823-8.

8. Goldacre MJ, Kurina LM, Seagroatt V, Yeates D. Abortion and breast cancer: a case-control record linkage study. *J Epidemiol Community Health* 2001; 55: 336-7.

9. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. *Cad Saúde Pública* 2003; 20(2): 362-71.

10. Blakely T, Atkinson J, Kiro C, Baiklock A, D'Souza A. Child mortality, socioeconomic position, and one-parent families: independent associations and variation by age and cause of death. *Int J Epidemiol* 2003; 32: 410-8.

11. Grundy E, Mayer D, Young H, Sloggett A. Living arrangements and place of death of older people with cancer in England and Wales: a record linkage study. *Br J Cancer* 2004; 91(5): 907-12.

12. Van Den Brabdt PA, Schouten L, Gold-Bohm RA, Dorant E, Hunen PHM. Development of a record linkage protocol for use in the Dutch Cancer Registry for epidemiological research. *Int J Epidemiol* 1990; 19: 553-8.

13. Almeida MF & Jorge MHPM. O uso da técnica de "Linkage" de sistemas de informação em estudos de coorte sobre mortalidade neonatal. *Rev Saúde Pública* 1996; 30(2): 141-7.

14. Horm J. *Linkage of the National Health Interview Survey with the National Death Index: methodological and analytic issues.* 1996. Disponível em htpp://www.cpc.unc.edu/pubs/paa_papers/1996/horm.html [Acessado em 26de janeiro de 2007].

15. Jones ME, Swerdlow AJ, Gill LE, Goldacre MJ. Pre-natal and early risk factors for childhood onset diabetes mellitus: a Record linkage study. *Int J Epidemiol* 1998; 27: 444-9.

16. Brewster DH, Stockton DL, Dobbie R, Bull D & Beral V. Risk of breast cancer after miscarriage or induced abortion: a Scottish record linkage case-control study. *J Epidemiol Community Health* 2005; 59(4): 283-7.

17. Holman CD, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* 1999; 23: 453-9.

18. Blakely T & Salmond C. Probabilistic record linkage and a method to calculate positive predictive value. *Int J Epidemiol* 2002; 31: 1246-52.

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

20. Newcombe HB & Kennedy, JM. Record Linkage making Maximum Use of the Discrimination Power of Identifying Information. *Commun ACM* 1962; 5: 563-6.

21. Fellegi IP & Sunter AB: A Theory for Record Linkage. *J Am Stat Assoc* 1969; 64: 1183-210.

22. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *J Am Stat Assoc* 1989; 84: 414-20.

23. Winkler WE. *Advanced Methods for Record Linkage. Technical Report.* Washington, DC: Statistical Research Division, U.S. Bureau of the Census; 1994. Disponível em http://www.census.gov/srd/www/byyear.html [Acessado em 12 de abril de 2007].

24. Portela MC, Schramm JMA, Pepe VLE, Noronha MF, Pinto CAM, Cianeli MP. Algoritmo para a composição de dados por internação a partir do sistema de informações hospitalares do sistema único de saúde (SIH/SUS) – Composição de dados por internação a partir do SIH/SUS. *Cad Saúde Pública* 1997; 13(4): 771-4.

25. 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.

26. Ravikumar P & Cohen WW. *A hierarchical graphical model for record linkag*e. Center for Automated Learning and Discovery, School of Computer Science, Carnegie Mellon University; 1969.

27. 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.

28. Machado CJ & Hill K. Determinantes da mortalidade neonatal e pós-neonatal no município de São Paulo. *Rev Bras Epidemiol* 2003; 6(4): 345-58.

29. Machado CJ & Hill K. Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem. *Cad Saúde Pública* 2004; 20(4): 915-25.

30. Kato I, Toniolo O, Koening KL, Kahn A, Schymura M, Zeleniuch-Jacquotte A. Comparison of Active and Cancer Registry-based Follow-up for Breast Cancer in a Prospective Cohort Study. Am J Epidemiol 1999; 149(4): 372-8.

31. Morais Neto OL & Barros MBA. Fatores de risco para mortalidade neonatal e pós-neonatal na Região Centro-Oeste do Brasil: linkage entre bancos de dados de nascidos vivos e óbitos infantis. *Cad Saúde Pública* 2000; 16(20): 477-85.

32. Churches T & Lim K. Using linkage to measure trends in breast cancer surgery. *NSW Public Health Bull* 2001; 12(4):105-11.

33. Saraceni V & Leal MC. Avaliação da efetividade das campanhas para eliminação da sífilis congênita na redução da morbi-mortalidade perinatal. Município do Rio de Janeiro, 1999-2000. *Cad Saúde Pública* 2003; 19(5): 1341-9.

34. Coeli CM, Coutinho ES, Veras RP. O desafio da aplicação da metodologia de captura-recaptura na vigilância do diabetes mellitus em idosos: lições de uma experiência no Brasil. Cad Saúde Pública 2004; 20(6): 1709-20.

35. Cavalcante MS, Ramos Junior AN, Pontes LRSK. Relacionamento de sistemas de informação em saúde: uma estratégia para otimizar a vigilância das gestantes infectadas pelo HIV. *Epidemiol Serv Saúde* 2005; 14(2): 127-33.

36. Brum L & Kupek E. Record linkage and capture-recapture estimates for underreporting of human leptospirosis in a Brazilian health district. *Braz J Infect Dis* 2005; 9(6): 515-20.

37. Sasson D, Silveira DP, Santos IS, Machado JP, Souza, SM, Mendes S. Diferenças no perfil de mortalidade da população brasileira e da população beneficiária de planos de saúde. In: *Brasil. Saúde Brasil 2006: uma análise da situação de saúde no Brasil*. Brasília: Ministério da Saúde; 2006. p. 105-207.

38. Brasil. Ministério do Planejamento, Orçamento e Gestão. Instituto Brasileiro de Geografia e Estatística. Ministério da Saúde. Pesquisa Nacional por Amostra de Domicílios: Acesso e Utilização de Serviços de Saúde, 2003. Rio de Janeiro, 2005. 167 páginas.

39. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise de Situação de Saúde. *Legislação Relativa aos Sistemas de Informações sobre Mortalidade (SIM) e sobre Nascidos Vivos* (SINASC). Brasília, 2005.

40. Brasil. Ministério da Saúde. Secretaria Executiva. Datasus. *Indicadores e Dados Básicos: Brasil 2005 – IDB 2005*. Disponível em http://tabnet.datasus.gov.br/cgi/idb2005/matriz.htm#cober [Acessado em 12 de abril de 2007].

Recebido em: 22/06/07
Versão final reapresentada em: 16/10/07
Aprovado em: 04/12/07