A construção da Base Nacional de Dados em Terapia Renal Substitutiva (TRS) centrada no indivíduo: relacionamento dos registros de óbitos pelo subsistema de Autorização de Procedimentos de Alta Complexidade (Apac/SIA/SUS) e pelo Sistema de Informações sobre Mortalidade (SIM) – Brasil, 2000-2004 *

Building the National Database on Renal Replacement Therapy Focused on the Individual: Probabilistic Record Linkage of Death Registries at the High Complexity Procedures Authorization Subsystem (Apac/SIA/SUS) and at the Mortality Information System (SIM) – Brazil, 2000-2004

Odilon Vanni de Queiroz
Mestrando em Saúde Pública pela Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

Augusto Afonso Guerra Júnior
Doutorando em Saúde Pública pela Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

Carla Jorge Machado
Departamento de Demografia, Faculdade de Ciências Econômicas, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

Eli Lola Gurgel Andrade
Departamento de Medicina Preventiva e Social, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

Wagner Meira Júnior
Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

Francisco de Assis Acúrcio
Departamento de Farmácia Social, Faculdade de Farmácia, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

Walter dos Santos Filho
Mestrando em Ciência da Computação pela Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

Mariângela Leal Cherchiglia
Departamento de Medicina Preventiva e Social, Faculdade de Medicina, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil

* Projeto financiado com recursos do Fundo Nacional de Saúde do Ministério da Saúde – FNS/MS –, UK Department for International Development, Organização das Nações Unidas para a Educação, a Ciência e a Cultura – UNESCO – e Conselho Nacional de Desenvolvimento Científico e Tecnológico do Ministério da Ciência e Tecnologia – CNPq/MCT.

Endereço para correspondência:
Av. Professor Alfredo Balena, 190, 7º Andar, Sala 706, Santa Efigênia, Belo Horizonte-MG, Brasil. CEP: 30130-100
E-mail: odilonvanni@gmail.com

Resumo

O relacionamento de registros vem sendo utilizado para integrar sistemas de informações em saúde. Neste trabalho, foram relacionados os registros de duas bases de dados entre 2000 e 2004: a Base Nacional de Dados em Terapia Renal Substitutiva (TRS), construída a partir dos dados do subsistema de Autorização de Procedimentos Ambulatoriais de Alta Complexidade (Apac) do Sistema de Informações Ambulatoriais do Sistema Único de Saúde (SIA/SUS); e o Sistema de Informações sobre Mortalidade (SIM). O objetivo do estudo foi comparar e complementar as informações de mortalidade da base TRS com informações do SIM. Os 176.773 registros da base TRS foram relacionados com 4.636.197 registros do SIM em três etapas, uma determinística e duas probabilísticas. Obteve-se uma concordância de 97,3% entre os pares julgados corretos, quando avaliados por dois revisores. O estudo demonstra as potencialidades da utilização do subsistema Apac/SIA/SUS, ainda pouco explorado, que, integrado a outros sistemas de informações em saúde, permite a organização da informação por paciente.

Palavras-chave: registro médico coordenado; sistemas de informações; registros de mortalidade; terapia renal substitutiva.

Summary

Record linkage has been used to integrate healthcare information systems. In this descriptive study in Brazil, records, from 2000 to 2004, of a National Database on Renal Replacement Therapy (TRS) *– built from the data available at the High Complexity Procedures Authorization Subsystem* (Apac) *of the Outpatient Information System/National Health System* (SIA/SUS) *– were linked to data available at the Mortality Information System* (SIM) *in order to compare and complement mortality information on both* TRS *and* SIM*. The records of 176,773 patients available at* TRS *were linked with 4,636,197 records available at* SIM*. The process has consisted of three stages, one deterministic and two probabilistic. The match of 97.3% of records from both systems found by two clerical reviewers (who agreed completely on their evaluation) shows the potential use of* Apac *– a yet little used system – when integrated to other health information systems to help organize information per patient.*

Key words*: medical record linkage; information systems; mortality registries; renal replacement therapy.*

Introdução

Os sistemas de informações disponíveis no Sistema Único de Saúde (SUS) são estratégicos na definição de prioridades e formulação de políticas de saúde. Entre tais sistemas, destacam-se: o Sistema de Informações Ambulatoriais (SIA/SUS), que contém dados da produção nacional de atendimentos em nível ambulatorial; o Sistema de Informações Hospitalares (SIH/SUS), com dados da produção nacional de atendimentos no âmbito hospitalar; o Sistema de Informações sobre Mortalidade (SIM), com informações de base populacional sobre mortalidade; e o Sistema de Informações sobre Nascidos Vivos (Sinasc), com os registros dos nascidos vivos no país.[1,2] Apesar do esforço para se criar uma identificação única de cada usuário do SUS, por meio do Cartão Nacional de Saúde, esses sistemas ainda não funcionam de forma integrada, tendo, inclusive, gestores diferentes. Cada sistema de informações abrange apenas uma dimensão específica do cuidado ou evento relativo à saúde e não permite que os indivíduos sejam facilmente rastreados em sua trajetória no SUS.[3,4]

> *A coincidência – ou complementariedade – das informações em dois sistemas distintos serve como evidência de sua confiabilidade.*

O relacionamento de registros, ou *record linkage*, representa uma alternativa para integrar dados dos sistemas de informações em saúde, ampliando o escopo de perguntas a serem respondidas, além de contribuir para a melhoria da qualidade dos dados registrados e permitir o seguimento longitudinal da assistência ao paciente.[5,6] Tal procedimento permite encontrar registros diferentes de uma mesma entidade em bases de dados distintas, ou identificar registros duplicados em uma mesma base de dados, podendo ser determinístico ou probabilístico.[2,7] O relacionamento determinístico considera como equivalentes os registros que forem concordantes (considerados iguais) em uma determinada chave (conjunto de identificadores). É indicado para situações em que exista uma chave unívoca entre os registros, como por exemplo, o cadastro de pessoa física (CPF). Em sua ausência, a tarefa é mais complexa. Pode-se utilizar uma combinação de campos, considerando-se equivalentes os registros que apresentam, por exemplo, datas de nascimento e nomes idênticos. Nestes casos, o relacionamento probabilístico é mais indicado, pois classifica pares de registros em prováveis, improváveis e duvidosos, levando-se em consideração as possibilidades de erros de preenchimento, grafia ou ocorrência de homônimos.[8]

A utilização de bancos de dados para analisar o padrão e os resultados do cuidado aos pacientes submetidos à terapia renal substitutiva (TRS) tem sido bastante encorajada. O subsistema de Autorização de Procedimentos Ambulatoriais de Alta Complexidade (custo) – Apac –, integrante do Sistema de Informações Ambulatoriais – SIA/SUS – é reconhecido como a maior fonte de dados sobre as TRS no Brasil, pelas informações epidemiológicas disponíveis e pela possibilidade de acompanhamento de séries históricas que ele permite.[9] Com o objetivo de realizar uma análise situacional das TRS, foi então construída, a partir do banco de dados administrativos do subsistema Apac/SIA/SUS, uma Base Nacional de Dados em TRS.[10]

A informação sobre a ocorrência do óbito, originalmente presente na base TRS, era oriunda somente do subsistema Apac/SIA/SUS. Neste subsistema, a informação depende da notificação dos óbitos pelos prestadores de serviços, que, eventualmente, podem não estar cientes desses óbitos. Além disso, essa informação no subsistema Apac/SIA/SUS é incompleta, pois não apresenta a causa do óbito.[11] Nesse sentido, o Sistema de Informações sobre Mortalidade – SIM –, baseado nas informações das declarações de óbito – DO – em âmbito nacional, pode ser de grande auxílio como fonte complementar às informações de óbito na base TRS.[12] A premissa básica é a seguinte: a coincidência – ou complementariedade – das informações em dois sistemas distintos serviria como evidência de sua confiabilidade. A base de dados nacional do SIM é gerada e administrada pela Secretaria de Vigilância em Saúde do Ministério da Saúde – SVS/MS – em parceria com o Departamento de Informática do SUS – Datasus.[13]

Este trabalho faz parte do projeto de pesquisa 'Avaliação Econômico-Epidemiológica das Terapias Renais Substitutivas no Brasil', e tem como objetivo relacionar os registros das bases TRS e SIM, descrevendo detalhadamente o procedimento. Outrossim, procurou-se

avaliar a confiabilidade dos pares considerados corretos pela inspeção manual.

Metodologia

O relacionamento probabilístico de registros alcançou sua formalização teórica e matemática com o trabalho de Fellegi e Sunter,[14] baseado na contribuição pioneira de Newcombe e colaboradores.[15] Os registros são comparados em pares e, posteriormente, classificados em prováveis, improváveis ou duvidosos. Esta classificação é feita com base em pesos de concordância e discordância, para cada identificador, definidos a partir da probabilidade condicional de concordância de cada identificador em pares verdadeiros (**m**), e na probabilidade condicional de concordância do identificador em pares falsos (**u**). Estes pesos podem assumir valores no intervalo de zero (inclusive) a 1 (inclusive).

No caso de concordância, a razão entre **m** e **u** é utilizada para decidir quais registros seriam considerados pares verdadeiros. E no caso de discordância, a razão entre **(1-m)** e **(1-u)** é utilizada na decisão de quais registros seriam considerados pares falsos. Usualmente, utiliza-se $\mathbf{log_2(m/u)}$ e $\mathbf{log_2[(1-m)/(1-u)]}$ como o peso do pareamento em caso de concordância e em caso de discordância, respectivamente. O escore final de cada par é resultado da soma dos pesos para cada identificador. Idealmente, um identificador adequado para o propósito do relacionamento probabilístico deve ter o valor de **m** próximo a 1 e o de **u** próximo a zero.[5]

Uma alternativa aos pesos de concordância utilizados, no caso dos identificadores possuírem distribuição de freqüências muito desigual, é utilizar um recurso chamado 'tabela de freqüência', em que o peso de concordância é o logaritmo na base 2 do inverso da freqüência relativa de cada categoria, ou valor, atribuída ao identificador.[16] Ou seja, o peso de concordância é definido pela função $\mathbf{F(x) = log_2[1/p(x)]}$, onde **p(x)** é a probabilidade de a variável assumir o valor **x**. Essa técnica se baseia no pressuposto de que valores mais raros de um identificador apresentam maior poder de discriminação, comparativamente aos mais freqüentes.[15] Por exemplo, se dois registros são concordantes quanto ao primeiro nome, essa concordância tem um peso maior para determinar que se trata de um mesmo indivíduo, no caso de um nome raro como 'Odilon'. No caso de um nome comum, como 'João', o peso de concordância atribuído deve ser menor.

Uma vez computado o escore para cada par (a soma dos pesos individuais dos identificadores), é gerado um gráfico da distribuição de freqüência dos pares segundo o escore obtido. A distribuição esperada dos escores é bimodal: os pares distribuídos em torno da primeira moda são os pares improváveis (de escores mais baixos); e os distribuídos em torno da segunda moda, os pares prováveis (com escores mais elevados). Os valores intermediários, compreendidos entre essas duas distribuições, são denominados pares duvidosos por não ser evidente a qual distribuição pertencem.

Operacionalmente, o relacionamento de registros consiste em três processos distintos: (1) padronização; (2) blocagem; e (3) *linkagem* de registros.[17]

A padronização dos registros é a primeira etapa do processo. Herzog, Scheuren e Winkler[18] subdividem essa etapa, também chamada de limpeza, em (i) padronização e (ii) divisão dos identificadores em termos (*parsing*). Seu objetivo é tornar tão grande quanto possível a probabilidade, pelo relacionamento, de campos equivalentes serem identificados como tais. A padronização consiste na codificação dos campos dos arquivos de dados em formato comum, para comparação, de forma que essa codificação seja consistente. Compreende, ainda, a eliminação de entradas fora de escopo e a verificação da integridade das bases. A divisão em termos consiste na subdivisão das variáveis, de forma a serem mais facilmente comparadas em um procedimento automático, via computador: por exemplo, a subdivisão de endereços em 'logradouro', 'número' e 'complemento'; ou a subdivisão de nomes em 'nome' e 'sobrenome'.

Para reduzir o custo computacional da comparação de todos os possíveis pares – que vem a ser o custo do produto cartesiano dos registros das bases comparadas –, utilizam-se técnicas de blocagem que permitem tão-somente a comparação de pares com maior probabilidade de equivalência. Segundo a tradição, o processo consiste na criação de partições dos arquivos, de tal maneira a serem comparados apenas os registros com um ou mais campos coincidentes entre as bases.[3] A terceira etapa – linkagem de registros – compreende o cômputo dos escores para cada par, em que são aplicados os pesos obtidos para cada variável, conforme já descrito.

Fonte dos dados

A Base Nacional de Dados em TRS foi construída a partir de registros identificados do subsistema Apac/SIA/SUS no período de 1° de novembro de 1999 a 31 de junho de 2005. A aplicação da técnica de relacionamento probabilístico permitiu a geração de um cadastro único de pacientes em TRS no Brasil.[10] A base TRS inclui informações para 176.773 pacientes que iniciaram alguma modalidade de TRS entre 2000 e 2004: variáveis demográficas (sexo, idade, Município, região de residência), clínicas [diagnóstico de causa de insuficiência renal crônica à entrada do paciente no sistema, segundo a Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde – Décima Revisão (CID-10)], de modalidade de tratamento (hemodiálise, diálise peritoneal e transplante renal], tempo de tratamento) e de resultados (óbito, continuidade de tratamento ou perda de seguimento); e variáveis relativas a gastos. A informação de óbitos era coletada, originalmente, pelo campo 'Motivo de cobrança' da Apac/SIA/SUS, cujos códigos 5.4, 9.1, 9.2 e 9,3 representam ocorrências relacionadas ao óbito.[11]

As bases de dados identificadas do Apac/SIA/SUS e do SIM foram obtidas junto ao Departamento de Economia da Saúde (DES), da Secretaria de Ciência, Tecnologia e Insumos Estratégicos do Ministério da Saúde (SCTIE/MS), um importante parceiro institucional no desenvolvimento do Projeto TRS. Para utilização das bases, firmou-se termo de compromisso e responsabilidade entre o DES/SCTIE/MS e o Grupo de Pesquisa em Economia da Saúde da Universidade Federal de Minas Gerais (GPES/UFMG). O *software* utilizado foi o Sistema Gerenciador de Banco de Dados MySQL,[19] versão 5.0. Por intermédio de uma rotina em linguagem SQL (Structured Query Language), foram realizados os processos de padronização, blocagem e linkagem. A rotina encontra-se disponível, mediante requisição encaminhada ao autor principal deste manuscrito.

As duas bases de dados utilizadas possuíam alguns identificadores comuns, os quais possibilitaram o relacionamento: 'nome completo do indivíduo'; 'nome completo da mãe'; 'sexo'; 'data de nascimento'; 'unidade da federação (UF) de nascimento'; e 'código IBGE do Município de residência', definido pela Fundação Instituto Brasileiro de Geografia e Estatística (IBGE). Além destes, a data do óbito no sistema SIM foi comparada à data de referência da última Apac do paciente na base TRS. Para tanto, partiu-se do princípio de que a insuficiência renal crônica é uma doença irreversível e, por conseguinte, os pacientes em TRS tenderiam a continuar sob tratamento até a data de seu óbito. Mesmo os pacientes submetidos a transplante renal permaneceriam em acompanhamento, recebendo medicamentos imunossupressores durante toda sua vida, o que é registrado na Apac. Sendo assim, é bastante razoável supor que a data de referência da última Apac do indivíduo seja próxima à data do eventual óbito.

A base SIM foi inspecionada para cada Estado, ano a ano, para verificar a completude dos dados. Cabe observar que, embora o propósito do presente estudo não seja o relacionamento anual entre as bases TRS e SIM, a observância de seu comportamento ao longo do tempo possibilitou identificar possíveis distorções nos números de registros esperados. De fato, observou-se ausência de dados de identificação nas declarações de óbito para alguns Estados brasileiros, referentes aos anos de 2000 e 2001. As UF mais acometidas por essa ausência foram São Paulo, Minas Gerais e Santa Catarina, conforme demonstra a Tabela 1. Cabe ressaltar: tal fato não representa sub-registro do SIM mas ausência de dados de identificação nas declarações de óbito no banco de dados disponibilizado, o que impossibilitou o relacionamento de registros.

Etapa de padronização dos dados

A padronização e a limpeza dos dados constituem a etapa mais trabalhosa e crítica do processo, por uma série de problemas de consistência e integridade das duas bases. Nessa fase, realizou-se uma busca por inconsistências que pudessem dificultar o relacionamento, tais como erros de preenchimento, dados incompletos ou ausentes. A freqüência de campos incompletos influi diretamente na probabilidade de obtenção de pares, especialmente quando se trata de bases de dados dependentes de poucos identificadores para seu pareamento. Para cada identificador, estabeleceu-se uma formatação que fosse comum entre as duas bases de dados, bem como um conjunto de valores válidos. Os valores não pertencentes a esse conjunto foram convertidos para 'NULO' e suas freqüências, discriminadas por identificador, são apresentadas na Tabela 2. A base TRS teve, ao todo, 7.065 registros (4,0%) com algum identificador convertido para 'NULO'. Essa informação, contudo, não reflete diretamente as freqüências de

inconsistências nos dados do subsistema Apac/SIA/SUS, uma vez que houve tratamento desses dados na construção da base. O SIM, entretanto, apresentou uma freqüência significativa de registros para os quais essa conversão foi efetuada: ao todo, 1.172.430 (25,3%). Muito embora esse percentual possa parecer elevado, é mister destacar que mais da metade desses registros identificados (595.753) era de declarações de óbito referentes aos anos 2000 e 2001, para os quais alguns Estados foram ausentes quanto a esses registros (Tabela 1). Ao se analisar apenas o período de 2002 a 2004, esse percentual cai para 19,2%. Ademais, a maioria desses registros teve apenas um dos campos ausente. Observou-se que as variáveis com maior freqüência de 'NULO' no SIM foram o 'nome completo da mãe' e a 'unidade da federação de nascimento': 13,2% e 14,4%, respectivamente. No caso da UF de nascimento, os registros apresentavam apenas a classificação do indivíduo como 'brasileiro'.

Os identificadores 'nome completo do indivíduo' e 'nome completo da mãe' receberam tratamento similar, qual seja, foram convertidos para letra maiúscula, tiveram retirados os acentos ortográficos e excluídos os espaços duplos, os espaços antes ou após o nome, além de quaisquer caracteres que não fossem letra (de A a Z). Uma dificuldade encontrada com relação a esses identificadores foi a utilização de uma grande diversidade de valores para refletir a ausência de informação, como 'NÃO IDENTIFICADO', 'INDIGENTE', 'NÃO INFORMADO'. Alguns desses valores apareciam com grande freqüência nas duas bases de dados, o que representaria um fator de viés para o resultado do relacionamento, uma vez que os pares em que esses valores co-ocorressem receberiam escores referentes à concordância no nome. Efetuou-se, então, uma busca exaustiva dos valores, que foram substituídos por 'NULO'.

O passo seguinte foi a subdivisão do nome do indivíduo e do nome da mãe, em primeiro nome, último nome e nome do meio. O primeiro e o último nome representaram, respectivamente, a primeira e a última palavra do nome constante do registro; e o nome do meio, tudo o que estivesse entre o primeiro e último nome, excluindo-se os conectivos 'de', 'do', 'da', 'dos' e 'das'.

A data de nascimento foi codificada em números inteiros de oito casas decimais: as quatro primeiras para o ano, as duas seguintes para o mês e as duas últimas para o dia (aaaammdd). Foram excluídos os valores cujo ano estivesse fora do intervalo de 1850 a 2004, o mês fora do intervalo 1 a 12, ou o dia fora do intervalo 1° a 31. O ano de nascimento posterior a 2004 foi excluído: a base TRS compreende dados de pacientes que iniciaram a TRS até 31 de dezembro de 2004.

A variável 'sexo' foi codificada como 'M' ou 'F', excluindo-se quaisquer outros valores. Com relação à UF de nascimento, para os brasileiros, manteve-se a sigla própria de cada unidade da federação; e para representar os estrangeiros, atribuiu-se o valor 99, uma vez que estes eram representados por códigos numéricos específicos em cada base de dados.

Tabela 1 - Declarações de óbito do SIM [a] que apresentavam identificação por ano, para os Estados de São Paulo (SP), Minas Gerais (MG) e Santa Catarina (SC). Brasil, 2000 a 2004

Ano	SP	MG	SC
2000	573	7.074	2.737
2001	957	96.284	27.814
2002	235.221	96.908	28.358
2003	238.039	104.234	29.330
2004	234.214	102.887	29.378

Fonte: Universidade Federal de Minas Gerais, Grupo de Pesquisa em Economia da Saúde – Base Nacional de Dados em Terapia Renal Substitutiva –; e Ministério da Saúde – Sistema de Informações sobre Mortalidade (2000-2004).

a) SIM: Sistema de Informações sobre Mortalidade

Devido à grande variação no código IBGE do Município de residência para um mesmo indivíduo, foram comparados somente os dois primeiros dígitos do código, que identificam a UF de residência. A tabela de Municípios brasileiros disponibilizada pelo Datasus possui códigos compreendidos no intervalo 110000 a 530010.[20] Foram excluídos os códigos não contidos nesse intervalo.

Finalmente, procedeu-se ao tratamento dos identificadores 'data do óbito' (SIM) e 'data de referência da última Apac' na base TRS, para cada indivíduo (Datref). Os campos foram codificados como números inteiros de seis casas decimais: as quatro iniciais para o ano e as duas finais para o mês (aaaamm). Excluíram-se valores para os quais o ano não era compreendido entre 2000 e 2004, ou aqueles para os quais o registro do mês não estivesse representado entre 1 e 12.

Foram inseridos dois campos adicionais em cada base, exclusivamente para a aplicação do algoritmo de codificação fonética Soundex, no primeiro e último nome do indivíduo. O algoritmo Soundex retorna um código que representa a interpretação fonética para as palavras analisadas. Como o algoritmo foi desenvolvido tomando por referência o idioma inglês, foram necessárias algumas adaptações para nomes brasileiros que apresentam variações de grafia na primeira sílaba, para um mesmo som, conforme descrito por Coeli e Camargo Jr.[3]

Relacionamento determinístico

O relacionamento determinístico tem por objetivo diminuir o número de pares a serem comparados nos segmentos posteriores. Nesse caso, foram considerados como pertencentes ao mesmo indivíduo os pares de registros das duas bases cuja correspondência fosse exata, após padronização, nos seguintes identificadores: primeiro e último nome do indivíduo; primeiro e último nome da mãe; data de nascimento; sexo; e Município de residência.

Etapa de blocagem

A blocagem constituiu-se de dois segmentos: no primeiro, utilizou-se o código Soundex para o primeiro e último nome do indivíduo; e no segundo, a equivalência exata da data de nascimento, sexo e UF de residência.

Por convenção, quando se utilizam estratégias de blocagem seriadas, elas são aplicadas ordenadamente, da mais restrita para a menos restrita, e os registros relacionados na etapa anterior são excluídos da etapa

Tabela 2 - Freqüência de valores ausentes ou inconsistentes na Base Nacional de Dados em TRS [a] e na base SIM.[b] Brasil, 2000 a 2004

Identificador	Base TRS [a] n=176.773		Base SIM [b] n=4.636.197	
	Absoluta	Relativa (%)	Absoluta	Relativa (%)
Sexo	–	–	2.869	0,06
Data de nascimento	129	0,07	96.956	2,09
UF [c] de nascimento	–	–	609.689	13,15
Município de residência	–	–	722	0,02
Maior data de referência ou data do óbito	–	–	14	0,00
Nome do indivíduo	1	0,00	456.895	9,85 [d]
Nome da mãe	6.950	3,93	665.724	14,36 [d]
Qualquer identificador acima	7.065	4,00	1.172.430	25,29 [d]

Fonte: Universidade Federal de Minas Gerais, Grupo de Pesquisa em Economia da Saúde – Base Nacional de Dados em Terapia Renal Substitutiva –; e Ministério da Saúde – Sistema de Informações sobre Mortalidade (2000-2004).

a) TRS: terapia renal substitutiva

b) SIM: Sistema de Informações sobre Mortalidade

c) UF: unidade da federação

d) Esses percentuais diminuem para 2,99%, 8,26% e 19,20%, respectivamente para os valores destacados, quando considerado somente o período de 2002 a 2004.

subseqüente.[3] Quando se exclui um registro que tenha sido relacionado em uma primeira estratégia de blocagem, entretanto, descarta-se a possibilidade de que ele venha a formar outro par com um escore superior, na estratégia seguinte. Por exemplo:

Suponha-se que seja necessário relacionar dois conjuntos de dados,

$$\mathbf{A = \{a_1, a_2, a_3\}} \text{ e } \mathbf{B = \{b_1, b_2, b_3\}}$$

e tome-se, como pressuposto, que $\mathbf{a_1}$ seja o par verdadeiro de $\mathbf{b_1}$. Porém, $\mathbf{a_2}$ é parecido com $\mathbf{b_1}$ o bastante para terem a mesma chave na primeira estratégia de blocagem e escore acima de um ponto do corte definido (assim, esquematicamente, a regra define que $\mathbf{a_2=b_1}$). Por algum erro de preenchimento, $\mathbf{a_1}$ e $\mathbf{b_1}$ (que, de fato, pertencem à mesma pessoa) têm chaves de blocagem diferentes na primeira estratégia, embora fossem comparáveis na segunda. Ao se excluir $\mathbf{a_1}$ ou $\mathbf{b_1}$ após a primeira estratégia de blocagem, elimina-se a possibilidade de que ambos sejam comparados na segunda estratégia, cuja regra de blocagem é diferente (esquematicamente, pela regra, $\mathbf{a_1 \neq b_1}$). Para evitar problemas desse tipo, os registros comparados na primeira estratégia de blocagem foram novamente comparados na segunda. E os pares por esta gerados, unidos e classificados.

As estratégias de blocagem foram definidas com o intuito de reduzir o quantitativo de pares a um número viável para relacionamento, sem repetir qualquer campo na chave de blocagem dessas etapas. Dessa forma, se um par não fosse comparado devido a erro de preenchimento de um dos campos de uma etapa, ainda poderia ser comparado em outra.

Etapa de relacionamento dos registros (linkagem)

O nome do indivíduo e o nome da mãe (primeiro nome, nome do meio e último nome) foram comparados utilizando-se o algoritmo de comparação aproximada de Jaro-Winkler.[21] O algoritmo retorna um valor entre zero e 1, em que 1 representa concordância exata. No presente estudo, adotou-se como ponto de corte o valor de 0,9 para 'primeiro e último nomes' e o valor de 0,8 para 'nome do meio', procurando-se minimizar erros na definição de pares corretos. A data do óbito e a maior data de referência no subsistema Apac/SIA/SUS foram consideradas equivalentes quando a diferença entre elas fosse menor ou igual a três meses. As demais variáveis foram comparadas de forma exata.

Quanto ao critério adotado na definição dos pesos de concordância-discordância para as variáveis 'sexo', 'data de nascimento', 'nome do meio' e 'data do óbito', foram utilizados os valores de **m** e **u** conforme a técnica usual. Quando algum dos valores a serem comparados resultou nulo, o peso adotado foi a média aritmética dos pesos de concordância e discordância, ou seja: o peso de concordância foi $\mathbf{\log_2(m/u)}$; o de discordância, $\mathbf{\log_2[(1-m)/(1-u)]}$; e quando um dos valores apresentou-se 'NULO',

$$\mathbf{\{\log_2(m/u)+\log_2[(1-m)/(1-u)]\}/2}.$$

Para o primeiro e último nomes do indivíduo e da mãe, assim como o Estado de nascimento e residência, foi calculado o peso de concordância com base na tabela de freqüência, pelo fato de seus valores possuírem uma distribuição de freqüência bastante desigual. O cálculo dos pesos com base em tabelas de freqüência seguiu estratégias distintas, para as variáveis comparadas de forma exata e para as variáveis comparadas de forma aproximada. Lembre-se de que as variáveis comparadas de forma exata foram 'unidade da federação de nascimento' e 'Estado de residência'; e as comparadas de forma aproximada (pelo algoritmo de Jaro-Winkler), 'primeiro e último nomes' do indivíduo e da mãe.

As tabelas de freqüência utilizadas foram geradas a partir da própria base de dados. Para os identificadores comparados de forma exata, as tabelas de freqüência foram obtidas da base TRS, uma vez que seriam consultadas somente em caso de concordância exata (valores idênticos nas duas bases). Para os identificadores comparados de forma aproximada, havia a possibilidade de valores serem considerados equivalentes, porém não idênticos. Neste caso, os dois valores poderiam ter freqüências distintas. Para esses campos, foi gerada uma tabela de freqüência para cada uma das duas bases: quando ocorreu equivalência exata dos valores comparados, adotou-se o escore calculado pela tabela de freqüência na base TRS. Quando esses valores não foram idênticos, embora semelhantes o suficiente como para serem considerados equivalentes, atribuiu-se o peso de concordância do valor mais freqüente em sua base de origem. Ou seja, em caso de valores não idênticos, com a finalidade de escolher qual seria selecionado na tabela de freqüência, optou-se por aquele que incorresse em menor peso de concordância. Esta conduta conservadora considerou a possibilidade de que o

valor menos freqüente pudesse ser um erro de grafia do mais freqüente. Por exemplo: um determinado par possuía, após padronização, o primeiro nome na base TRS como 'CONCEIAO' e, no SIM, como 'CONCEICAO'. O resultado da comparação aproximada de Jaro-Winkler para os dois nomes foi de 0,98. Portanto, os dois nomes foram considerados equivalentes **mas não idênticos**. O peso de concordância calculado por meio da freqüência relativa da grafia 'CONCEIAO' na base TRS foi de 12,75, enquanto o do nome 'CONCEICAO' no SIM foi de 5,43. Assim, atribuiu-se o peso de concordância de 5,43 para esse identificador.

A Tabela 3 descreve os pesos de concordância e de discordância, e o peso atribuído para valores 'NULO', além dos valores de **m** e **u** utilizados para cada identificador.

Confiabilidade (mensurada para pares corretos)

Para avaliar a confiabilidade dos pares considerados verdadeiros pelo relacionamento, eles foram inspecionados manualmente e classificados como corretos ou incorretos por dois revisores – autores deste artigo – que trabalharam de forma independente. A estatística Kappa foi utilizada para avaliar a concordância entre os dois revisores.[22-24]

Essa estatística não foi adotada para avaliar a concordância dos revisores com o relacionamento, uma vez que o objetivo desta revisão não é comparar o método de relacionamento automático de registros com o procedimento manual e sim obter uma estimativa da qualidade da informação de óbito imputada pelo relacionamento probabilístico. Deste modo, não

Tabela 3 - Valores de 'm' e 'u' e respectivos pesos de concordância e discordância para o relacionamento dos registros da Base Nacional de Dados em TRS[a] e da base SIM.[b] Brasil, 2000 a 2004

Identificadores	m	u	1-m	1-u	Concordância	Discordância	Não declarado
Primeiro nome	0,90	0,01	0,10	0,99	3,44-17,84[d]	-3,31	1,59
Último nome	0,85	0,02	0,15	0,98	3,10-17,83[d]	-2,71	1,35
Primeiro nome da mãe	0,75	0,05	0,25	0,95	2,15-17,83[d]	-1,93	0,99
Ultimo nome da mãe	0,70	0,05	0,30	0,95	3,43-17,83[d]	-1,66	1,07
Data do óbito	0,92	0,02	0,08	0,98	5,52	-3,61	0,95
Data de nascimento	0,91	0,01	0,09	0,99	6,51	-3,46	1,52
Estado de residência	0,95	0,08	0,05	0,92	1,91-9,98[d]	-4,20	-0,32
UF[c] de nascimento	0,90	0,09	0,10	0,91	2,18-10,50[d]	-3,19	0,07
Sexo	0,98	0,51	0,02	0,49	0,94	-4,61	-1,84
Nome do meio	0,68	0,09	0,32	0,91	2,92	-1,51	0,70
Nome do meio da mãe	0,60	0,15	0,40	0,85	2,00	-1,09	0,46

Fonte: Universidade Federal de Minas Gerais, Grupo de Pesquisa em Economia da Saúde – Base Nacional de Dados em Terapia Renal Substitutiva –; e Ministério da Saúde – Sistema de Informações sobre Mortalidade (2000-2004).

a) TRS: terapia renal substitutiva

b) SIM: Sistema de Informações sobre Mortalidade

c) UF: unidade da federação

d) O peso de concordância para esses identificadores não foi utilizado porque foi calculado pela tabela de freqüência. Observa-se, então, a amplitude de variação dos escores obtidos com base nas tabelas de freqüência.

foram inspecionados pares considerados falsos pelo relacionamento automático.

Considerações éticas

O projeto de pesquisa 'Avaliação Econômico-Epidemiológica das Terapias Renais Substitutivas no Brasil' foi aprovado pela Comissão de Ética em Pesquisa da Universidade Federal de Minas Gerais (UFMG) (Parecer ETIC n° 397/ 2004).

Resultados

A Base Nacional de Dados em TRS possui 176.773 registros; e a base SIM, 4.636.197. O número de pares gerados em cada segmento de blocagem é apresentado na Tabela 4. Nos segmentos de blocagem 2 e 3, o tempo de processamento foi de 6 horas, aproximadamente.

Para o relacionamento determinístico, gerou-se uma tabela com 14.818 pares, que representaram registros de óbitos para os pacientes da base TRS. Para os segmentos 1 e 2 do relacionamento probabilístico, os pares encontrados foram avaliados como verdadeiros, falsos ou duvidosos, uma vez que as tabelas geradas para esses pares continham o escore obtido da comparação entre todos os pares e atendiam, portanto, aos critérios das respectivas blocagens.

Para o tratamento dos pares gerados no relacionamento probabilístico, desenhou-se uma tabela com os pares de maior escore para cada paciente da base TRS comparado nessas etapas, 'Tabela Maior Escore', ou 'Tabela ME', revelando-se um total de 235.167 pares formados. Em seguida, obteve-se a distribuição do logaritmo neperiano de freqüências dos pares da Tabela ME, segundo o escore (Figura 1). Adotou-se a escala logarítmica por sua capacidade de representar grandes variações de freqüência em um espaço menor e sua utilidade no trabalho com dados que cobrem uma extensa gama de valores. O logaritmo reduz a representação a uma escala mais facilmente visível – e manejável –, o que permite estabelecer a relação percentual entre os valores.

A distribuição do logaritmo de freqüências mostrou seu maior valor em torno do escore 15. Não se obteve a distribuição teórica bimodal; porém, a curva não se apresentou como uma normal "bem comportada", em forma de sino, revelando um platô que abrangeu pares do escore 30 ao 60, aproximadamente. Foram inspecionados, manualmente, pares com escore entre 25 (ponto a partir do qual a freqüência começou a declinar mais intensamente, indicando a possibilidade do início de uma distribuição de pares corretos) e 40. A partir do valor de 29,9, a proporção de pares verdadeiros mostrou ser superior à de pares falsos. Optou-se, então, por adotar esse valor como ponto de corte, a partir do qual poder-se-ia classificar um par como verdadeiro.

Os pares cuja maior data de referência no subsistema Apac/SIA/SUS fosse igual ou superior a março de 2005 foram considerados falsos, ainda que apresentassem escore acima do ponto de corte. Esse critério adicional revelou-se necessário, uma vez que foram comparadas as declarações de óbito até dezembro de 2004, não sendo razoável que um indivíduo continuasse a ter registros de Apac após a data do eventual óbito. Manteve-se, contudo, a mesma tolerância de três meses adotada na comparação da maior data de referência

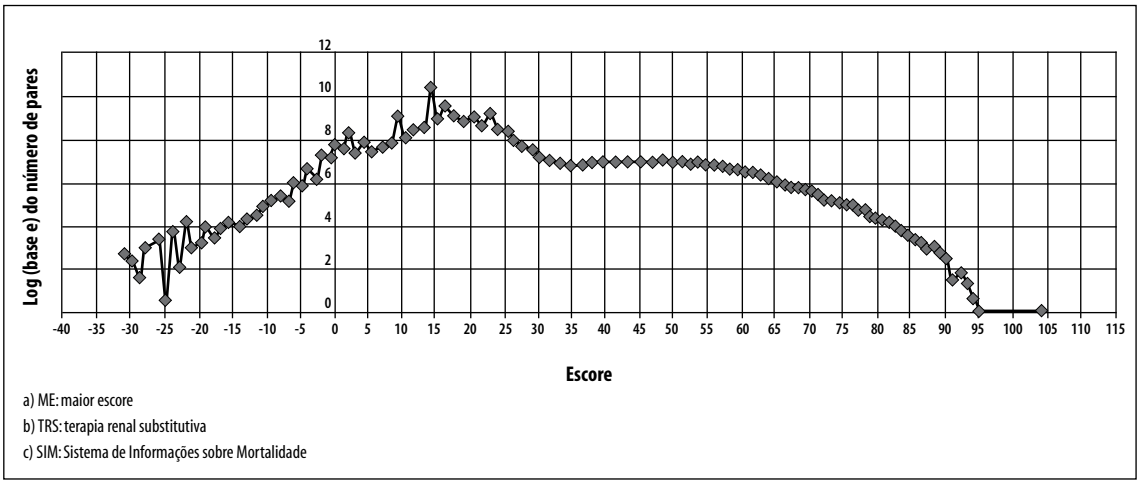
Tabela 4 - Número de pares gerados por segmento do relacionamento dos registros entre a Base Nacional de Dados em TRS [a] e a base SIM.[b] Brasil, 2000 a 2004

Segmento	Número de registros		Pares gerados
	Base TRS [a]	Base SIM [b]	
1 (determinístico)	176.773	4.636.197	14.818
2 (probabilístico)	161.955	4.636.197	523.077.601
3 (probabilístico)	161.955	4.636.197	10.220.137

Fonte: Universidade Federal de Minas Gerais, Grupo de Pesquisa em Economia da Saúde – Base Nacional de Dados em Terapia Renal Substitutiva –; e Ministério da Saúde – Sistema de Informações sobre Mortalidade (2000-2004).

a) TRS: terapia renal substitutiva

b) SIM: Sistema de Informações sobre Mortalidade



a) ME: maior escore
b) TRS: terapia renal substitutiva
c) SIM: Sistema de Informações sobre Mortalidade

Figura 1 - Distribuição do logaritmo natural da freqüência dos pares na Tabela ME [a] dos escores referentes ao relacionamento dos registros da Base Nacional de Dados em TRS [b] e da base SIM.[c] Brasil, 2000 a 2004

(Apac/SIA/SUS) com a da data do óbito (SIM). A inspeção manual permitiu verificar que os pares com data de referência no subsistema Apac/SIA/SUS posterior a março de 2005, não obstante o peso acima do ponto de corte, efetivamente eram falsos pares.

Estabelecidos os critérios para considerar um par como verdadeiro, segundo o princípio de que cada paciente deveria ter apenas uma declaração de óbito, gerou-se uma nova tabela a partir da Tabela ME: **Tabela ME_nova**. Desta tabela, constavam os indivíduos da base TRS presentes na Tabela ME que atendessem aos seguintes critérios: (i) escore acima do ponto de corte, (ii) maior data de referência no subsistema Apac/SIA/SUS, desde que anterior a março de 2005, e (iii) com apenas uma declaração de óbito.

Quanto aos pares encontrados mediante relacionamento determinístico, não foi necessário estabelecer ponto de corte. Identificaram-se, porém, 24 pacientes – distribuídos em 11 UF – com dois registros de óbito. Com o propósito de selecionar uma única declaração de óbito para esses indivíduos, comparou-se o campo "nome do meio", o qual ainda não havia sido utilizado: os indivíduos que possuíssem apenas uma declaração de óbito seriam, finalmente, incluídos na Tabela ME_nova, consolidando-se os resultados obtidos. Cumprido esse procedimento, a Tabela ME_nova passou a registrar 52.048 óbitos dos pacientes da Base Nacional de Dados em TRS que atenderam aos critérios de pares verdadeiros.

Concluído o relacionamento, os pares considerados válidos foram classificados em decis, a partir da distribuição de freqüências dos seus escores. De cada decil, retirou-se uma amostra de trinta pares, verificados manualmente por dois revisores independentes, justamente autores do presente estudo (OVQ e MLC), para serem classificados como verdadeiros ou falsos. Os dois revisores obtiveram total concordância na avaliação (Kappa=1,0). Foram encontrados cinco pares falsos no primeiro decil (16,7%) e três no segundo (10,0%). Nos decis subseqüentes, todos os pares foram classificados como verdadeiros. A proporção de pares considerados corretos por ambos os revisores foi de 97,3%, sobre um total de 300 pares inspecionados manualmente.

No relacionamento entre a Base Nacional de Dados em TRS e a base nacional do SIM, identificaram-se, entre 2000 e 2004, 52.048 óbitos no SIM e 45.203 no subsistema Apac/SIA/SUS (campo Motcob); 34.158 óbitos encontravam-se em ambas as bases. Em média, para o período, 75,6% dos óbitos registrados na Apac foram confirmados no SIM: 34.158/45.203. Nos anos 2000 e 2001, esse percentual foi de 54,5%, bastante inferior ao percentual médio (87,9%) dos três anos seguintes. O percentual inferior nos dois primeiros anos pode ser justificado pela falta de registros de identificação no SIM (Tabela 1). Não constituem objeto desta análise as diferenças nas informações relativas a óbito encontradas entre as duas bases de dados.

Discussão

O relacionamento de registros vem assumindo grande importância no cenário da Saúde Pública. As necessidades dos gestores e pesquisadores da área da Saúde, associadas à forma como foram estruturados os sistemas de informações em saúde no Brasil, determinam a necessidade de integrar dados desses sistemas, independentemente de apresentarem identificador unívoco. Ainda são raros os estudos envolvendo o subsistema Apac/SIA/SUS,[7] de modo que integrar os dados desse subsistema e os do sistema SIM foi de grande importância e utilidade para os autores deste trabalho. Além de possibilitar maior conhecimento da estrutura, potencialidades e deficiências do subsistema Apac/SIA/SUS, o estudo permitiu confirmar a informação de óbito, quando presente nesse subsistema, pelos dados constantes no sistema SIM. Para estudos futuros, a tarefa de integração de informações possibilitará o aproveitamento das informações conjugadas para esses dois bancos de dados, referentes à data e causa do óbito.

> *Este trabalho, ao relacionar bases de dados administrativos e epidemiológicos do SUS, abre caminho para novos estudos epidemiológicos, econômicos e avaliativos dos serviços de saúde.*

A limpeza e padronização dos dados demonstraram ser esta etapa a mais importante e trabalhosa do processo, dada a grande freqüência de dados inconsistentes, incompletos ou com erros de grafia. Os bancos de dados administrativos não foram projetados especificamente para fins de pesquisa e suas informações não se caracterizam pela alta qualidade exigida para essa finalidade.[1,2,4,7,24] O Sistema de Informações sobre Mortalidade, particularmente, possuía 25,3% dos registros com alguma informação inconsistente ou ausente. Essa particularidade do SIM obriga que o *software* de relacionamento a ser utilizado permita a identificação e tratamento, de forma diferenciada, dos valores *missing* ou ausentes, uma vez que, para efeito de relacionamento de registros, não parece razoável que a concordância – ou discordância – entre valores dessa natureza corrobore a declaração de um par como verdadeiro ou falso.[25,26]

Ao longo de cinco anos, este estudo de abrangência nacional compreendeu o relacionamento de 176.773 registros da Base Nacional de Dados em TRS e 4.636.197 registros do SIM. O relacionamento de bases tão grandes, mediante a técnica probabilística, é bastante desafiador e raramente encontrado na literatura. Sob esse aspecto, a utilização do *software* MySQL mostrou ser uma alternativa robusta, ademais com a versatilidade necessária para o tratamento dos valores ausentes ou inválidos nas bases. Entre outras vantagens, o MySQL possui código aberto e funciona em inúmeros sistemas operacionais, tais como Windows e Linux, entre outros. É portável, ou seja, funciona na maioria dos computadores, com excelente desempenho e estabilidade. É importante acrescentar, no entanto, que esse *software* não foi criado especificamente para o relacionamento de registros. Trata-se de um instrumento gerenciador de banco de dados e seu uso requer a codificação dos procedimentos desejados em linguagem SQL.[19] A utilização desse recurso tem precedentes na literatura: Drumond, França e Machado, ao utilizarem uma rotina em linguagem SQL para o relacionamento de registros do Sistema de Informações Hospitalares – SIH/SUS – com o Sistema de Informações sobre Nascidos Vivos – Sinasc –, obtiveram bons resultados.[27]

Outros *softwares* de relacionamento de registros apresentam a vantagem de não exigir codificação de rotinas para seu uso. Entre eles, estes autores destacam três, de distribuição gratuita: o Reclink, *software* desenvolvido por pesquisadores brasileiros;[28] o Febrl, criado pela Universidade Nacional Australiana;[25] e o Link Plus, desenvolvido e adotado pelos Centers for Disease Control and Prevention (CDC) de Atlanta-GA, Estados Unidos da América.[26] Em etapas preliminares deste trabalho, foram realizados testes com o Reclink (versão 2.1.7.200), haja vista esse aplicativo ser bastante utilizado por pesquisadores no Brasil; e com o Febrl (versão 0.3), por apresentar grande variedade de recursos e ser de código aberto. Optou-se, contudo, pela não-utilização de ambos: no caso do Reclink, por não implementar alguns dos recursos utilizados neste trabalho, tais como cálculo do peso de concordância por tabela de freqüência e tratamento diferenciado para valores *missing*; e do Febrl, por não ter apresen-

tado o desempenho necessário para o relacionamento de registros em número tão grande quanto o utilizado por este trabalho, principalmente devido ao excessivo consumo de memória.

Alguns resultados preliminares deste relacionamento encorajam estudos futuros. O SIM confirmou óbitos notificados na base Apac/SIA/SUS, para os anos de 2002 a 2004, em proporção elevada (87,9%). Como o SIM tem cobertura estimada de 82% dos óbitos ocorridos no país, com variações regionais,[12] o fato de não se ter confirmado 100% dos óbitos é um resultado coerente e demonstra que a técnica de relacionamento aplicada foi satisfatória. Ademais, entre os pares considerados verdadeiros pelo relacionamento, 97,3% foram ratificados pela inspeção manual, proporção esta bastante satisfatória.

Apesar das dificuldades encontradas para a consecução deste trabalho, os autores deste estudo conseguiram relacionar, satisfatoriamente, bases administrativas e epidemiológicas do SUS. Sua utilização abre caminho para novos estudos epidemiológicos, econômicos e de avaliação dos serviços de saúde, de grande importância para a formulação de políticas específicas e melhoria da qualidade da atenção prestada aos pacientes submetidos às terapias de substituição renal no país.

Referências

1. Carvalho DM. Grandes sistemas nacionais de informação em saúde: revisão e discussão da situação atual. Informe Epidemiológico do SUS 1997;5:7-46.
2. Pinheiro RS, Camargo Jr KR, Coeli CM. Relacionamento de bases de dados em saúde. Cadernos de Saúde Coletiva 2006;14:195-196.
3. Coeli CM, Camargo Jr KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Revista Brasileira de Epidemiologia 2002;5:185-196.
4. Mendes ACG, Silva Junior JB, Medeiros KR, Lyra TM, Melo Filho DA, Sá DA. Avaliação do Sistema de Informações Hospitalares – SIH/SUS como fonte complementar na vigilância e monitoramento de doenças de notificação compulsória. Informe Epidemiológico do SUS 2000;9:67-86.
5. Teixeira CLS, Block KV, Klein CH, Coeli CM. Método de relacionamento de bancos de dados do Sistema de Informação sobre Mortalidade (SIM) e das autorizações de internação hospitalar (BDAIH) no Sistema Único de Saúde (SUS), na investigação de óbitos de causa mal-definida no Estado do Rio de Janeiro, Brasil, 1998. Epidemiologia e Serviços de Saúde 2006;5:47-57.
6. Veras CMT, Martins MAS. Confiabilidade dos dados nos formulários de Autorização de Internação Hospitalar (AIH) – Rio de Janeiro, Brasil. Cadernos de Saúde Pública 1994;10:339-355.
7. 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. Cadernos de Saúde Coletiva 2006;14:197-224.
8. Scheuren F. Linking health records: human rights concerns. Proceedings of an International Workshop and Exposition: Record Linkage Techniques; 1997. Washington (DC): National Academy Press; 1999.
9. Ministério da Saúde. Política Nacional ao Portador de Doença Renal. Brasília: MS; 2004.
10. Cherchiglia ML, Guerra Júnior AA, Andrade EIG, Machado CJ, Acúrcio FA, Meira Júnior 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. Revista Brasileira de Estudos de População 2007; 24:163-167.
11. Ministério da Saúde. Sistema de Informações Ambulatoriais do SUS - SIA/SUS: manual de orientações técnicas. Brasília: MS; 2006.
12. Gomes Jr SCS, Almeida RT. Comparação do registro da produção ambulatorial em oncologia no Sistema Único de Saúde. Cadernos de Saúde Pública 2006;22:141-150.
13. Laurenti R, Mello Jorge MHP, Gotlieb SLD. A confiabilidade dos dados de mortalidade e morbidade por doenças crônicas não-transmissíveis. Ciência & Saúde Coletiva 2004;9:909-920.
14. Fellegi IP, Sunter A. A theory of record linkage. Journal of the American Statistical Association 1969;64:1183-1210.

15. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science 1959;130:954-959.

16. Conn L, Bishop G. Exploring methods for creating a longitudinal census dataset. Australian Bureau of Statistics; 2005.

17. Camargo Jr KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. Cadernos de Saúde Pública 2000;16:439-447.

18. Herzog TN, Sheuren FJ, Winkler WE. Data quality and record linkage techiniques. Springer; 2007.

19. MySQL. The world's most popular open source database [database on the Internet]. Sweden: MySQL. c2995. [cited 2008 Jul. 14]. Available from: http://www.mysql.com.

20. Ministério da Saúde. Datasus. Bem vindo ao repositório de tabelas corporativas [dados na Internet]. Brasília: MS [acesso 14 jul. 2008]. Disponível em: http://repositorio.datasus.gov.br.

21. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods. American Statistical Association; 1990. p. 354-359.

22. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.

23. Escosteguy CC, Portela MC, Medronho RA, Vasconcellos MTL. O Sistema de Informações Hospitalares e a assistência ao infarto agudo do miocárdio. Revista de Saúde Pública 2002; 36:491-499.

24. Mathias TAF, Soboll MLMS. Confiabilidade de diagnósticos nos formulários de Autorização de Internação Hospitalar. Revista de Saúde Pública 1998;32:526-532.

25. Christen P. Febrl - A Freely Available Record Linkage System with a Graphical User Interface. Second Australasian Workshop on Health Data and Knowledge Management (HDKM 2008); 2008; Wollongong, NSW, Australia. Australian Computer Science Communications; 2008.

26. Centers for Disease Control and Prevention. National Program of Cancer Registries. Link Plus [homepage on the Internet]. Atlanta: CDC [cited 2008 Jul. 14]. Available from: http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm.

27. Drumond EdF, França EB, Machado CJ. SIH-SUS e Sinasc: utilização do método probabilístico para relacionamento de dados. Cadernos de Saúde Coletiva 2006;14:251-264.

28. Reclink. Relacionamento probabilístico de registros [dados na Internet] [acesso 14 jul. 2008]. Disponível em: http://paginas.terra.com.br/educacao/kencamargo/RecLink.html

Recebido em 18/03/2008
Aprovado em 12/09/2008