Relacionamento probabilístico de registros: uma aplicação na área de morbidade materna grave (*near miss*) e mortalidade materna

Probabilistic record linkage: an application to severe maternal morbidity (near miss) and maternal mortality

Maria Helena de Sousa [1]
José Guilherme Cecatti [1]
Ellen Hardy [1]
Suzanne Jacob Serruya [2]

[1] Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brasil.
[2] Departamento de Ciência e Tecnologia, Ministério da Saúde, Brasília, Brasil.

Correspondência
J. G. Cecatti
Departamento de Tocoginecologia, Faculdade de Ciências Médicas, Universidade Estadual de Campinas.
Rua Alexander Fleming 101, Campinas, SP 13083-881, Brasil.
cecatti@unicamp.br

Abstract

This article presents an experience with record linkage from the Brazilian Hospital Information System (SIH) with the Mortality Information System (SIM), and the SIH with itself, applied to severe maternal morbidity (near miss) and maternal mortality. This was an empirical study using Brazilian data for the state capitals and Federal District in 2002. For the two linkages separately applied in each capital, a three simple step blocking strategy was established, plus related multiple steps and also two clerical review strategies. From the total number of true pairs found after the two linkages, simple steps failed to find fewer than 8%, while the multiple step strategy failed to find only 0.7%. This approach allowed exploring the issue of severe maternal morbidity and mortality in these databases. The number of pairs found and reviewed under the multiple steps strategy was lower than the sum of pairs obtained with the three simple steps, and fewer pairs were lost. However, for the record linkage of the SIH with itself, both strategies are suggested.

Information Systems; Maternal Mortality; Mosidity

Introdução

Em 1946, Dunn [1] publicou, em um periódico de saúde pública, um artigo em que o termo *record linkage* foi abordado pela primeira vez. O autor fez uma analogia entre a história de vida de uma pessoa e um livro, cujo início e fim corresponderiam, respectivamente, ao nascimento e ao óbito da mesma: *record linkage* seria o processo necessário para reunir as páginas deste livro em um único volume.

Alguns anos depois, Newcombe et al. [2] publicaram um artigo sobre o tema, apresentando resultados de um estudo em que aplicaram *record linkage* a dados de registros vitais. Abordaram pela primeira vez o tópico de cálculo das probabilidades e o logaritmo de base 2 das mesmas, como parte da teoria de informação. Uma década depois, Fellegi & Sunter [3] produziram uma extensa teoria sobre *record linkage*.

Apesar da base teórica envolvida em *record linkage*, destaca-se que o enfoque principal refere-se à aplicação puramente prática, quando se tem, por exemplo, registros de dois bancos de dados (arquivos): um na área da saúde e outro na área de dados vitais, cujas informações necessitam ser confrontadas, com o objetivo de se estabelecer a correspondência ou não dos pares de registros, cada um deles proveniente de um dos arquivos.

Em geral o *record linkage* é realizado com o intuito de obter apenas um banco de dados,

combinando as informações parciais existentes em cada arquivo original para obtenção de um arquivo único completo, ou com o objetivo de construir e manter um arquivo mestre com informações atualizadas [3].

Tal processo seria relativamente simples e direto se houvesse um campo que identificasse inequivocamente cada registro como, por exemplo, um número de identificação único por indivíduo ou vários campos que, juntos, permitissem identificar os registros de uma mesma pessoa. Nessa situação, os arquivos poderiam ser confrontados utilizando-se o chamado relacionamento determinístico [4]. Uma possibilidade futura nesse sentido para o Brasil, é a adoção do já recomendado registro único de saúde ou número SUS [5], a exemplo do que já acontece em países desenvolvidos como Suécia, Noruega e Dinamarca [4].

No Brasil, dada a inexistência desse campo único nos arquivos de saúde, a identificação direta é impossível de ser realizada. Dessa forma, faz-se necessário utilizar diversas informações menos específicas, procurando-se estabelecer o quão provável um par de registros corresponde a um mesmo indivíduo ou unidade de análise. Trata-se do chamado relacionamento probabilístico, anteriormente mencionado, que foi primeiramente abordado por Newcombe et al. [2] e posteriormente por Fellegi & Sunter [3].

Machado [4], em revisão sobre *record linkage* determinístico e probabilístico, com aplicação na área de saúde infantil, apresentou as linhas que diversos países têm seguido nestes dois tipos de relacionamento. O Brasil, apesar de ter sido referido como o único país em desenvolvimento que tem abordado o tema, ainda está em estágios iniciais de estudos.

A concreta utilização das fontes de dados do Ministério da Saúde brasileiro, por meio de seus diversos sistemas de informações de rotina (entre eles o de mortalidade, o de nascidos vivos e o de informações hospitalares), implicaria uma promissora área de pesquisa aplicada, incluindo a avaliação da qualidade destes sistemas. Entretanto, apesar da existência desses bancos de dados eletrônicos do sistema brasileiro de saúde, que estão facilmente disponíveis [6], os mesmos não apresentam informações de identificação dos registros, por questões de ética e sigilo. Portanto, campos como nome e endereço não são disponibilizados. Dessa forma, o método de relacionamento probabilístico somente poderia ser utilizado se informações de identificação pudessem ser obtidas.

Para operacionalizar essa necessidade, Camargo Jr. & Coeli [7] desenvolveram um programa chamado RecLink II, que executa o relacionamento probabilístico de registros, quando se tem diversos campos em comum para serem comparados. Esse programa é de domínio público, sendo obtido na Internet (http://paginas.terra.com.br/educacao/kencamargo/RecLink.html).

O objetivo do presente estudo foi apresentar uma experiência empírica de relacionamento do Sistema de Informações Hospitalares (SIH) com o Sistema de Informações sobre Mortalidade (SIM), e do SIH com ele próprio, utilizando-se diferentes critérios nas etapas de blocagem e de revisão manual, e simultaneamente verificar a possibilidade de explorar a morbidade materna grave (*near miss*) e mortalidade materna nestes bancos de dados de informações rotineiras em saúde disponíveis no Brasil.

Método

Tratou-se de um estudo empírico explorando as possibilidades metodológicas de relacionamento de bancos de dados para a extração de informações relativas a indivíduos comuns, com base em dados brasileiros do SIH e do SIM. Os arquivos SIH e SIM, com dados para as capitais de estados e o Distrito Federal em 2002, foram especialmente obtidos do Departamento de Informática do SUS (DATASUS) e da Secretaria de Vigilância em Saúde (SVS) do Ministério da Saúde, respectivamente. Esses arquivos continham informações que permitiram identificar as mulheres (como nome e data de nascimento) para fins de relacionamento dessas bases de dados. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Ciências Médicas da Universidade Estadual de Campinas (parecer nº. 147/2004). Foi resguardado o sigilo do nome de cada pessoa identificada.

Com base no SIH houve uma seleção inicial de registros de internações de mulheres de 10 a 49 anos de idade, seguida de uma seleção daqueles que apresentaram pelo menos um item indicativo de morbidade materna grave (*near miss*). Esta última seleção baseou-se em uma lista previamente elaborada, baseada fundamentalmente nos critérios internacionalmente recomendados e originários dos estudos de Mantel et al. [8] e de Waterstone et al. [9]. Para os dados do SIM houve a seleção de registros de mulheres de 10 a 49 anos de idade. Na seqüência, foram separados os registros das 26 capitais de estados e do Distrito Federal, para cada um dos dois sistemas.

Os campos principais utilizados para os relacionamentos foram o nome e a data de nascimento. O programa utilizado para o relacionamento probabilístico dos sistemas foi o RecLink II [7]. Esse programa é dividido em várias etapas seqüenciais: padronização dos bancos de dados;

relacionamento propriamente dito (subdividido em blocagem e pareamento); combinação dos arquivos e revisão manual. A primeira etapa faz-se necessária apenas uma vez, enquanto as demais são repetidas em vários passos, para diferentes chaves de blocagem, de acordo com o estabelecido pela avaliação subjetiva do pesquisador.

A padronização dos arquivos envolveu a transformação de todos os caracteres para letra maiúscula, a exclusão de acentos, bem como a eliminação das preposições "de", "da", "do", "dos", "das" e de sinais de pontuação erroneamente digitados. Também permitiu a subdivisão dos campos nome e data de nascimento, cujas partes foram utilizadas na etapa seguinte de blocagem e de pareamento.

Para os dois relacionamentos (SIH *vs.* SIM e SIH *vs.* SIH), aplicados separadamente em cada uma das capitais, estabeleceu-se estratégia de blocagem em três passos simples, bem como estratégia de blocagem em múltiplos passos relacionados. As chaves de blocagem utilizadas nos passos simples foram: (1) código fonético (*Soundex*) do primeiro nome formatado para blocagem – Pbloco; (2) código fonético (*Soundex*) do último nome formatado – Ubloco; e (3) ano de nascimento – Anonas. A estratégia de múltiplos passos relacionados apresentou um passo inicial, denominado Pi, com uma chave mais restrita envolvendo o código fonético do primeiro e último nomes e o ano de nascimento, ou seja, *Soundex*(Pbloco) + *Soundex*(Ubloco) + Anonas. Após esse passo inicial, foram utilizadas as três chaves isoladamente, em etapas sucessivas e dependentes das anteriores (com exclusão de pares já localizados anteriormente). Para o segundo relacionamento, do SIH com ele mesmo (para localização de reinternações de uma mesma pessoa), aplicaram-se mais dois passos extras (denominado P[E]), com o nome completo e endereço residencial como chaves de blocagem complementares e utilizando-se o arquivo original, sem exclusões.

Os fatores de ponderação de concordância e discordância, para o cálculo dos escores [3], são obtidos respectivamente pelas seguintes equações: $\log_2(m_i/u_i)$ e $\log_2[(1-m_i)/(1-u_i)]$, em que m_i é a probabilidade dos campos concordarem dado que se trata de par verdadeiro (equivalente à sensibilidade); u_i é a probabilidade dos campos concordarem dado que se trata de par falso (1-especificidade), $(1-m_i)$ refere-se ao complementar da sensibilidade e, finalmente, $(1-u_i)$ equivale à especificidade. Neste estudo, para a composição do escore total de cada par, foram utilizados valores sugeridos por Camargo Jr. & Coeli (http://paginas.terra.com.br/educacao/kencamargo/RecLink.html) para vários campos na etapa de pareamento. Esses campos, com o algoritmo de comparação, bem como os valores de m_i e u_i utilizados para o cálculo dos escores constam da Tabela 1.

Exemplificando o cálculo do escore total na situação de concordância completa do primeiro nome, último nome, inicial(is) do(s) nome(s) do meio e data de nascimento, o escore máximo será obtido por meio da soma apenas dos fatores de ponderação de concordância, ou seja, $\log_2(m_i/u_i)$. Portanto, nesse caso teremos a soma que resultará em aproximadamente 25,3: $\log_2(99/2) + \log_2(99/3) + \log_2(89/3) + \log_2(91/10) + \log_2(94/23) + \log_2(92/4)$, cujas parcelas correspondem, respectivamente, aos escores dos campos: primeiro nome, último nome, inicial(is) do(s) nome(s) do meio, dia, mês e ano de nascimento.

Após a combinação dos arquivos pelo RecLink II, chegou-se à etapa final de revisão manual dos pares combinados. Apesar de existirem

Tabela 1

Valores da sensibilidade, complementar da especificidade e proporção mínima de concordância (valores sugeridos por Camargo Jr. & Coeli; http://paginas.terra.com.br/educacao/kencamargo/RecLink.html), para os campos utilizados no pareamento dos bancos de dados.

Campo	Algoritmo de comparação %	Sensibilidade (mi) %	1-Especificidade (ui) %	Proporção mínima de concordância
Primeiro nome padronizado *	Aproximado	99	2	75
Último nome padronizado *	Aproximado	99	3	75
Iniciais do(s) nome(s) do meio	Exato	89	3	-
Dia do nascimento	Pela diferença	91	10	2
Mês de nascimento	Pela diferença	94	23	2
Ano de nascimento	Pela diferença	92	4	2

* Denominados Pbloco (primeiro nome) e Ubloco (último nome).

fórmulas para estimativa de valores limiares [3], como sua determinação não é simples e direta, decidiu-se neste estudo avaliar empiricamente os escores positivos mais altos, em detrimento daqueles mais baixos.

Com isso, a revisão manual inicial da junção SIH *vs.* SIM ocorreu para todos os pares cujos escores foram iguais ou maiores que 18; os nomes foram verificados (especialmente o primeiro e o segundo nomes); se houvesse dúvida, as datas de nascimento eram verificadas; se ainda necessário, a data de saída pelo SIH era confrontada com a data do óbito pelo SIM. Também foi feita uma revisão manual de escores iguais ou maiores que 1 e inferiores a 18, porém com seleção de pares que tinham datas de nascimento idênticas nos dois sistemas, mas com data de saída pelo SIH igual ou inferior à data do óbito pelo SIM, ou que tinham datas de nascimento diferentes, porém com data de saída pelo SIH igual à data do óbito pelo SIM; o principal campo verificado foi o nome (especialmente o primeiro e o segundo) e, se necessário, as datas de nascimento eram observadas, bem como as datas de saída (SIH) e de óbito (SIM), se as mesmas eram próximas.

A revisão manual inicial da junção SIH *vs.* SIH (o arquivo com ele mesmo), para localização de várias internações de uma mesma pessoa, ocorreu inicialmente para os pares cujos escores foram iguais ou maiores que 17, verificando-se principalmente os nomes e as datas de nascimento. Também foi feita uma revisão manual de registros cujos escores resultaram iguais ou maiores que 1 e inferiores a 17, porém com seleção de pares que tinham pelo menos dois dos três – dia, mês e ano de nascimento – iguais e a diferença entre a data de internação no segundo arquivo e a data de saída no primeiro arquivo não ultrapassasse trinta dias; ou que as datas de nascimento fossem idênticas nos dois sistemas, mas as datas de saída fossem diferentes; ou que tivessem datas de nascimento diferentes, mas com as datas de saída iguais nos dois arquivos. Para os dois passos extras, a faixa de escore para verificação manual ficou compreendida entre -5 e 20.

Utilizou-se análise descritiva simples, para os dois relacionamentos (SIH *vs.* SIM e SIH *vs.* SIH), e os programas computacionais usados foram o RecLink II e o SPSS versão 11.5 (SPSS Inc., Chicago, Estados Unidos).

Resultados

Para o primeiro relacionamento, do SIH com o SIM, e com estratégia de passos simples, houve a formação mínima de quase um milhão de pares sob a terceira estratégia de blocagem (passo 3), e a máxima de mais de dois milhões de pares, sob a segunda (passo 2). Do total de 151 pares verdadeiros, 4 (2,6%) não foram localizados no passo 1; 11 (7,3%) no passo 2 e; 6 (4%) no passo 3. A capital que apresentou maior número de pares formados e revisados manualmente, pelas três estratégias, foi São Paulo (Tabela 2).

Com a aplicação da estratégia de múltiplos passos relacionados, 132 (87,4%) dos 151 pares foram localizados no passo inicial, com a chave mais restrita, enquanto que 18 foram localizados em algum dos três passos seguintes, e apenas um par (0,7%), de Fortaleza (Estado do Ceará), não foi localizado após todo o processo múltiplo (Tabela 3).

No segundo *linkage*, do SIH com ele mesmo, para os 1.405 pares verdadeiros em todas as capitais, 45 (3,2%) não foram encontrados no passo 1; 87 (6,2%) no passo 2 e; 75 (5,3%) no passo 3. Novamente, a capital que apresentou o maior número de pares formados e revisados foi São Paulo (Tabela 4).

Após a aplicação dos múltiplos passos, verificou-se que 1.239 (88,2%) dos 1.405 pares verdadeiros foram encontrados no Pi; 137 (9,8%) em algum dos três passos seguintes; 19 (1,3%) nos dois passos extras e; apenas 10 (0,7%) não puderam ser localizados no processo completo (Tabela 5).

Discussão

O número de pares formados em cada um dos relacionamentos foi da ordem de 10^6, enquanto que o de pares revisados foi da ordem de 10^3. Enquanto as estratégias de passos simples perderam menos de 8% do total de pares verdadeiros, para os dois relacionamentos (SIH *vs.* SIM e SIH *vs.* SIH), a estratégia de múltiplos passos perdeu bem menos, ou seja, menos de 1%. Para cada um dos relacionamentos, o número de pares formados (e revisados) sob a estratégia completa de múltiplos passos foi inferior à soma dos pares formados (e revisados) nos três passos simples.

O tema de *record linkage* não é novo, pois a primeira publicação foi há mais de 60 anos [1]. Entretanto, apesar do desenvolvimento tecnológico das últimas décadas, com equipamentos cada vez mais rápidos e modernos, o elemento humano ainda acaba envolvido no processo de revisão manual dos pares, como responsável pela decisão final sobre o pareamento ou não. Apesar da subjetividade do ser humano e da possibilidade real de se errar, a máquina não o substitui nessa etapa do relacionamento. Um estudo recente abordou o problema da tomada de decisão em

Tabela 2

Número de pares formados e número de pares revisados para dois intervalos de escore segundo estratégia de blocagem, e número de pares verdadeiros e perdidos, para as 27 capitais brasileiras (*linkage* Sistema de Informações Hospitalares – SIH *vs.* Sistema de Informações sobre Mortalidade – SIM).

Capital/UF	Passo 1: *Soundex*(Pbloco)			Passo 2: *Soundex*(Ubloco)			Passo 3: Anonas			Número de pares verdadeiros	Número de pares perdidos		
	Pares formados	Pares revisados Escore		Pares formados	Pares revisados Escore		Pares formados	Pares revisados Escore			Passo		
		1|-18 *	≥ 18		1|-18 *	≥ 18		1|-18 *	≥ 18		1	2	3
Porto Velho/RO	3.371	0	2	3.722	0	2	1.643	0	2	1	0	0	0
Rio Branco/AC	5.158	1	1	3.123	0	1	1.277	1	1	1	0	0	0
Manaus/AM	9.392	2	1	9.172	0	1	5.327	2	1	2	0	1	0
Boa Vista/RR	119	0	0	250	0	0	139	0	0	0	-	-	-
Belém/PA	44.372	7	5	33.557	6	6	22.968	8	6	7	1	1	1
Macapá/AP	1.362	0	3	1.190	1	3	1.107	1	3	3	0	0	0
Palmas/TO	895	1	3	1.019	0	3	537	1	3	4	0	1	0
São Luís/MA	18.332	3	1	13.660	4	1	8.345	6	1	1	0	0	0
Teresina/PI	41.458	2	2	29.788	8	2	9.193	2	1	1	0	0	0
Fortaleza/CE	203.350	16	9	109.696	13	9	36.142	9	5	6	1	1	2
Natal/RN	7.139	1	3	10.374	1	3	2.140	0	3	3	0	0	0
João Pessoa/PB	2.747	0	3	4.144	1	3	946	1	3	3	0	0	0
Recife/PE	20.551	5	5	78.296	8	5	9.624	4	2	1	0	0	0
Maceió/AL	22.934	11	8	51.283	13	8	6.296	5	4	3	0	0	0
Aracaju/SE	3.766	4	4	12.590	3	4	1.503	2	2	4	0	2	0
Salvador/BA	56.670	11	13	234.567	13	12	63.584	12	13	10	1	1	0
Belo Horizonte/MG	24.101	1	10	35.973	2	10	21.249	3	10	10	0	0	0
Vitória/ES	310	0	2	661	0	2	415	0	2	1	0	0	0
Rio de Janeiro/RJ	184.391	20	15	388.463	28	16	168.688	30	10	11	0	0	0
São Paulo/SP	735.416	105	71	1.206.639	129	72	519.896	133	49	50	0	2	3
Curitiba/PR	9.933	1	3	13.293	0	3	14.782	3	3	3	0	0	0
Florianópolis/SC	221	0	0	219	0	0	325	0	0	0	-	-	-
Porto Alegre/RS	8.339	0	6	11.450	3	6	10.250	4	6	6	0	0	0
Campo Grande/MS	2.384	1	1	3.478	0	1	2.608	1	1	1	0	0	0
Cuiabá/MT	701	2	5	2.910	1	5	1.602	2	5	7	0	1	0
Goiânia/GO	2.862	0	6	7.148	0	7	3.104	1	7	7	1	0	0
Brasília/DF	72.324	11	6	88.863	11	5	42.694	7	6	5	0	1	0
Total	1.482.598	205	188	2.355.528	245	190	956.384	238	149	151	4	11	6

Unidades da Federação – AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceará; DF: Distrito Federal; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; RR: Roraima; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

* Datas de nascimento idênticas nos dois sistemas e data da saída pelo SIH ≤ data do óbito pelo SIM ou datas de nascimento diferentes nos dois sistemas e data da saída pelo SIH = data do óbito pelo SIM.

pares duvidosos [10], porém não havia o nome para ser considerado como campo de blocagem e/ou pareamento, fato não ocorrido neste estudo.

Uma primeira limitação deste estudo foi o fato de não termos avaliado a qualidade dos sistemas utilizados, devido à impossibilidade operacional deste processo. Sabe-se apenas que vários casos de morte materna declarada, existentes no SIM, não foram localizados no SIH, quando a base de referência foi esta última. Outra limitação refere-se à qualidade das informações de identificação existentes nos sistemas, entre elas o nome e o endereço. Neste estudo, o campo nome parece ter sido digitado de forma padronizada, pois, por exemplo, não se verificaram muitos casos de nomes abreviados e com erros de digitação.

Um estudo que considerou um banco de dados do SIH relacionado a um inquérito domici-

Tabela 3

Número de pares formados e número de pares revisados para dois intervalos de escore segundo estratégia de blocagem em múltiplos passos, e número de pares perdidos e localizados, para as 27 capitais brasileiras (*linkage* Sistema de Informações Hospitalares – SIH *vs.* Sistema de Informações sobre Mortalidade – SIM).

Capital/UF	Passo inicial – Pi *Soundex*(Pbloco) + *Soundex*(Ubloco) + Anonas			Três passos seguintes – P[3] P1: *Soundex*(Pbloco); P2: *Soundex*(Ubloco); P3: Anonas			Número de pares perdidos	Número de pares localizados	
	Pares formados	Pares revisados Escore		Pares formados	Pares revisados Escore			Passo	
		1|-18*	≥ 18		1|-18 *	≥ 18		Pi	P[3]
Porto Velho/RO	5	0	2	3.722	0	2	0	1	0
Rio Branco/AC	6	0	1	3.123	0	1	0	1	0
Manaus/AM	7	0	1	9.172	0	1	0	1	1
Boa Vista/RR	0	-	-	250	0	0	0	-	-
Belém/PA	35	0	5	33.557	6	6	0	4	3
Macapá/AP	4	0	3	1.190	1	3	0	3	0
Palmas/TO	3	0	3	1.019	0	3	0	3	1
São Luís/MA	10	2	1	13.660	4	1	0	1	0
Teresina/PI	34	0	1	29.788	8	2	0	1	0
Fortaleza/CE	166	0	5	109.696	13	9	1	4	1
Natal/RN	15	0	3	10.374	1	3	0	3	0
João Pessoa/PB	8	0	3	4.144	1	3	0	3	0
Recife/PE	66	0	2	78.296	8	5	0	1	0
Maceió/AL	81	1	4	51.283	13	8	0	3	0
Aracaju/SE	17	0	2	12.590	3	4	0	2	2
Salvador/BA	103	0	12	234.567	13	12	0	8	2
Belo Horizonte/MG	22	0	9	35.973	2	10	0	10	0
Vitória/ES	2	0	2	661	0	2	0	1	0
Rio de Janeiro/RJ	157	1	10	388.463	28	16	0	11	0
São Paulo/SP	820	5	48	1.206.639	129	72	0	45	5
Curitiba/PR	8	0	3	13.293	0	3	0	3	0
Florianópolis/SC	0	-	-	219	0	0	0	-	-
Porto Alegre/RS	8	0	6	11.450	3	6	0	6	0
Campo Grande/MS	4	0	1	3.478	0	1	0	1	0
Cuiabá/MT	6	1	5	2.910	1	5	0	6	1
Goiânia/GO	9	0	6	7.148	0	7	0	6	1
Brasília/DF	72	0	5	88.863	11	5	0	4	1
Total	1.668	10	143	4.717.161	647	123	1	132	18

Unidades da Federação – AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceará; DF: Distrito Federal; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; RR: Roraima; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

* Datas de nascimento idênticas nos dois sistemas e data da saída pelo SIH ≤ data do óbito pelo SIM ou datas de nascimento diferentes nos dois sistemas e data da saída pelo SIH = data do óbito pelo SIM.

liar, encontrou uma baixa proporção de registros identificados nos dois arquivos e os autores sugerem que uma das razões pode ter sido a questionável qualidade do preenchimento das informações utilizadas para a identificação [11].

As estratégias de blocagem em passos simples resultaram em muitos casos formados, com um mínimo de aproximadamente um milhão. Apesar dos recursos computacionais modernos, o tempo envolvido não deveria ser desprezado. O mesmo se aplica ao tempo gasto na tarefa de revisão manual, que neste estudo não foi pouco, embora não tenha sido computado.

Sabe-se que a utilização de blocagem implica otimização do processo de relacionamento de bancos de dados [7], e que a revisão manual é um

Tabela 4

Número de pares formados e número de pares revisados para dois intervalos de escore segundo estratégia de blocagem, e número de pares verdadeiros e perdidos, para as 27 capitais brasileiras (*linkage* Sistema de Informações Hospitalares – SIH *vs.* Sistema de Informações sobre Mortalidade – SIM).

Capital/UF	Passo 1: *Soundex*(Pbloco)			Passo 2: *Soundex*(Ubloco)			Passo 3: Anonas			Número de pares verdadeiros	Número de pares perdidos		
	Pares formados Refrec < Assrec	Pares revisados Escore 1|-17 *	≥ 17	Pares formados Refrec < Assrec	Pares revisados Escore 1|-17 *	≥ 17	Pares formados Refrec < Assrec	Pares revisados Escore 1|-17 *	≥ 17		Passo 1	2	3
Porto Velho/RO	1.798	1	5	4.026	2	5	3.319	2	3	5	0	0	1
Rio Branco/AC	4.995	6	20	5.265	2	20	5.032	6	18	23	0	3	2
Manaus/AM	2.512	1	2	3.542	4	2	4.841	5	1	2	0	0	1
Boa Vista/RR	56	0	3	127	0	3	160	0	2	3	0	0	1
Belém/PA	50.816	23	45	79.294	41	45	120.279	57	42	43	1	1	1
Macapá/AP	1.691	0	10	1.878	0	10	3.067	1	9	9	0	0	1
Palmas/TO	1.364	4	17	3.770	3	17	3.582	2	15	17	0	1	1
São Luís/MA	15.329	12	14	23.085	16	16	31.707	29	16	18	2	1	0
Teresina/PI	88.222	68	54	101.281	87	55	70.853	81	43	61	6	12	11
Fortaleza/CE	227.387	82	126	174.494	86	131	134.312	105	121	121	5	1	4
Natal/RN	5.329	5	32	12.544	4	34	5.528	5	29	32	2	0	4
João Pessoa/PB	482	0	4	1.630	2	4	798	2	4	4	0	0	0
Recife/PE	11.510	14	68	71.678	39	71	21.837	43	59	58	1	4	3
Maceió/AL	21.928	24	56	81.845	46	58	22.674	34	45	47	1	2	3
Aracaju/SE	3.532	4	10	20.308	17	10	5.897	16	7	8	0	2	0
Salvador/BA	60.038	46	77	352.998	168	81	199.887	226	58	59	1	6	3
Belo Horizonte/MG	14.776	13	66	35.530	12	65	41.231	22	65	69	1	5	2
Vitória/ES	175	0	5	518	1	5	724	1	5	5	0	0	0
Rio de Janeiro/RJ	88.883	91	183	272.798	136	186	231.024	239	160	190	14	19	25
São Paulo/SP	415.325	170	362	1.091.439	394	372	818.786	467	315	314	3	11	7
Curitiba/PR	9.870	7	85	17.682	4	85	34.262	19	83	87	0	4	1
Florianópolis/SC	147	0	5	181	0	5	383	0	5	5	0	0	0
Porto Alegre/RS	4.672	4	80	8.685	5	80	15.093	10	80	81	0	1	0
Campo Grande/MS	2.073	5	20	4.104	1	20	5.901	5	19	25	0	4	1
Cuiabá/MT	602	2	17	3.184	1	18	3.327	2	18	20	2	2	0
Goiânia/GO	1.030	1	17	3.220	1	18	3.603	2	18	20	2	1	0
Brasília/DF	80.381	37	78	170.475	71	82	146.632	88	73	79	4	7	3
Total	1.114.923	620	1.461	2.545.581	1.143	1.498	1.934.739	1.469	1.313	1.405	45	87	75

Unidades da Federação – AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceará; DF: Distrito Federal; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; RR: Roraima; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

Refrec: numeração seqüencial do primeiro arquivo; Assrec: numeração seqüencial do segundo arquivo.

Nota: Refrec e Assrec foram gerados pelo RecLink II (http://paginas.terra.com.br/educacao/kencamargo/RecLink.html).

* Pelo menos dois dos três: dia, mês e ano de nascimento iguais e data da internação SIHb-data da saída SIHa ≤ 30 dias ou datas de nascimento idênticas nos dois sistemas e data da saída pelo SIHa ≠ data da internação SIHb ou datas de nascimento diferentes e data da saída pelo SIHa = data da internação SIHb.

processo lento, trabalhoso, que depende da avaliação subjetiva do revisor, porém necessário nos casos duvidosos.

Segundo Coeli & Camargo Jr. [12], a estratégia de blocagem mais eficiente é a de múltiplos passos relacionados, o mesmo sendo verificado neste estudo. Entretanto, foi possível observar que na comparação do SIH com ele mesmo, para localização de reinternações de uma mesma mulher, os casos com mais de duas internações e cujos escores eram inferiores a 17 (ponto de corte utilizado) não puderam ser localizados sob a estratégia de múltiplos passos. Dessa forma, sugere-se que sejam aplicadas as duas estratégias (múltipla e simples) no caso de se relacionar um banco de dados com ele mesmo. Outro artigo derivado da

Tabela 5

Número de pares formados e número de pares revisados para dois intervalos de escore segundo estratégia de blocagem em múltiplos passos, e número de pares perdidos e localizados, para as 27 capitais brasileiras (*linkage* Sistema de Informações Hospitalares – SIH *vs.* Sistema de Informações sobre Mortalidade – SIM).

Capital/UF	Passo inicial – Pi *Soundex*(Pbloco) + *Soundex*(Ubloco) + Anonas			Três passos seguintes – P[3] P1: *Soundex*(Pbloco); P2: *Soundex*(Ubloco); P3: Anonas			Dois passos extras – P[E] Nome; endereço		Número de pares perdidos	Número de pares localizados		
	Pares formados Refrec < Assrec	Pares revisados Escore 1|-17 *	≥ 17	Pares formados Refrec < Assrec Escore < 25,3	Pares revisados Escore 1|-17 *	≥ 17	Pares formados Refrec < Assrec Escore < 25,3	Pares revisados Escore -5|-20 *		Passo Pi	P[3]	P[E]
Porto Velho/RO	7	1	3	8.847	2	3	62	1	0	4	1	0
Rio Branco/AC	27	0	18	14.175	11	2	46	3	0	18	5	0
Manaus/AM	6	1	1	10.779	10	1	150	0	0	1	1	0
Boa Vista/RR	2	0	2	321	0	1	5	2	0	2	1	0
Belém/PA	121	2	42	240.413	119	11	860	8	0	40	3	0
Macapá/AP	9	0	9	6.358	1	4	45	1	0	8	1	0
Palmas/TO	20	0	15	7.439	5	3	4	1	0	15	2	0
São Luís/MA	37	2	14	67.820	52	2	798	19	0	15	3	0
Teresina/PI	383	8	42	245.715	207	24	642	45	0	40	15	6
Fortaleza/CE	586	4	116	485.080	255	45	688	37	0	111	10	0
Natal/RN	41	0	27	19.745	11	11	56	4	0	26	6	0
João Pessoa/PB	8	0	4	2.759	0	0	18	0	0	4	0	0
Recife/PE	147	3	59	96.184	89	43	905	81	0	52	4	2
Maceió/AL	232	5	45	114.979	98	34	197	53	0	43	3	1
Aracaju/SE	31	1	7	29.089	31	12	223	23	1	6	1	0
Salvador/BA	302	6	56	589.580	413	63	1562	56	2	49	8	0
Belo Horizonte/MG	89	2	64	84.241	36	8	161	5	2	63	4	0
Vitória/ES	5	0	5	1.320	2	0	22	3	0	5	0	0
Rio de Janeiro/RJ	377	5	156	532.810	436	59	863	67	2	154	25	9
São Paulo/SP	1.426	13	312	2.132.563	932	163	1565	236	2	293	19	0
Curitiba/PR	91	0	83	54.071	27	4	233	7	0	83	4	0
Florianópolis/SC	5	0	5	659	1	0	2	0	0	5	0	0
Porto Alegre/RS	86	0	80	24.195	16	0	96	1	0	80	1	0
Campo Grande/MS	23	1	19	11.101	3	1	29	1	1	20	4	0
Cuiabá/MT	19	0	17	6.526	4	1	6	1	0	17	3	0
Goiânia/GO	20	0	17	7.196	2	4	4	2	0	17	3	0
Brasília/DF	219	2	70	376.905	176	31	223	27	0	68	10	1
Total	4.319	56	1.288	5.170.870	2.939	530	9.464	684	10	1.239	137	19

Unidades da Federação – AC: Acre; AL: Alagoas; AM: Amazonas; AP: Amapá; BA: Bahia; CE: Ceará; DF: Distrito Federal; ES: Espírito Santo; GO: Goiás; MA: Maranhão; MG: Minas Gerais; MS: Mato Grosso do Sul; MT: Mato Grosso; PA: Pará; PB: Paraíba; PE: Pernambuco; PI: Piauí; PR: Paraná; RJ: Rio de Janeiro; RN: Rio Grande do Norte; RO: Rondônia; RR: Roraima; RS: Rio Grande do Sul; SC: Santa Catarina; SE: Sergipe; SP: São Paulo; TO: Tocantins.

Refrec: numeração seqüencial do primeiro arquivo; Assrec: numeração seqüencial do segundo arquivo.

Nota: Refrec e Assrec foram gerados pelo RecLink II (http://paginas.terra.com.br/educacao/kencamargo/RecLink.html).

* Pelo menos dois dos três: dia, mês e ano de nascimento iguais e data da internação SIHb-data da saída SIHa ≤ 30 dias ou datas de nascimento idênticas nos dois sistemas e data da saída pelo SIHa ≠ data da internação SIHb ou datas de nascimento diferentes e data da saída pelo SIHa = data da internação SIHb.

exploração analítica desses mesmos bancos de dados mostra a utilização direta da estratégia de blocagem em múltiplos passos relacionados [13].

Apesar das limitações dos sistemas existentes, foi possível explorar o assunto de morbidade materna grave nos bancos de dados de informações rotineiras em saúde disponíveis no Brasil. Entretanto, convém salientar novamente que isso só foi possível porque os bancos de dados com informações de identificação foram espe-

cialmente fornecidos para este estudo, contendo estas informações não disponíveis normalmente. Embora o estudo tenha demonstrado a possibilidade da utilização desses bancos de dados para o estudo da morbidade materna grave e mortalidade materna, a dificuldade operacional encontrada permite a proposição da adoção de mecanismos de informação mais eficientes pelo sistema oficial de saúde. A proposta da utilização de um único número identificador do indivíduo no sistema seria razoável como estratégia para simplificar o processo e melhorar sua eficiência para fins de monitoramento.

Resumo

*Apresentar uma experiência de relacionamento do Sistema de Informações Hospitalares (SIH) com o Sistema de Informações sobre Mortalidade (SIM), e do SIH com ele próprio, aplicados na área de morbidade materna grave (*near miss*) e mortalidade materna. Trata-se de um estudo empírico, utilizando-se dados brasileiros das capitais de estados e do Distrito Federal em 2002. Para os dois relacionamentos, aplicados separadamente em cada uma das capitais, estabeleceu-se estratégia de blocagem em três passos simples, bem como a de múltiplos passos relacionados, e duas de revisão manual. Do total de pares verdadeiros dos dois relacionamentos, menos de 8% não puderam ser localizados pelos passos simples, enquanto que a estratégia de múltiplos passos deixou de localizar apenas 0,7%. Foi possível explorar o assunto de mortalidade e morbidade materna grave nos bancos de dados. O número de pares formados e revisados sob a estratégia de múltiplos passos foi inferior à soma dos pares nos três passos simples e, além disso, menos pares foram perdidos. Porém, para o relacionamento do SIH com ele próprio, sugerem-se as duas estratégias.*

Sistemas de Informação; Mortalidade Materna; Morbidade

Colaboradores

M. H. Sousa foi responsável pela implementação do estudo e análise estatística. Os quatro autores participaram do plano de análise. M. H. Sousa escreveu a primeira versão do artigo, complementada a seguir por J. G. Cecatti. Todos os autores discutiram, leram e aprovaram a versão final do artigo.

Agradecimentos

Este estudo foi parcialmente financiado pelo Human Reproduction Programme, World Health Organization, (projeto H9-181R862).

Referências

1. Dunn HL. Record linkage. Am J Public Health 1946; 36:1412-6.
2. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science 1959; 130:954-9.
3. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc 1969; 64:1183-210.
4. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. Cad Saúde Pública 2004; 20:362-71.
5. Ministério da Saúde. O SUS no seu município: garantindo saúde para todos. Brasília: Ministério da Saúde; 2004.
6. Fundação Nacional de Saúde/Departamento de Informática do SUS. Diretório de bases de dados. http://datasus.gov.br/dirbd/estrut.htm (acessado em 04/Abr/2003).
7. Camargo Jr. KR, Coeli CM. *Reclink*: aplicativo para o relacionamento de bases de dados, implementando o método *probabilistic record linkage*. Cad Saúde Pública 2000; 16:439-47.
8. Mantel GD, Buchmann E, Rees H, Pattinson RC. Severe acute maternal morbidity: a pilot study of a definition for a near-miss. Br J Obstet Gynaecol 1998; 105:985-90.
9. Waterstone M, Bewley S, Wolfe C. Incidence and predictors of severe obstetric morbidity: case-control study. BMJ 2001; 322:1089-93.
10. Machado CJ, Hill K. Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem. Cad Saúde Pública 2004; 20:915-25.
11. Coeli CM, Blais R, Costa MCE, Almeida LM. Probabilistic linkage in household survey on hospital care usage. Rev Saúde Pública 2003; 37:91-9.
12. Coeli CM, Camargo Jr. KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol 2002; 5:185-96.
13. Sousa MH, Cecatti JG, Hardy EE, Serruya SJ. Morte materna declarada e o relacionamento de sistemas de informações em saúde. Rev Saúde Pública 2007; 41:181-9.

Recebido em 18/Mai/2006
Versão final reapresentada em 14/Ago/2007
Aprovado em 13/Set/2007