# *Reclink*: aplicativo para o relacionamento de bases de dados, implementando o método *probabilistic record linkage*

## Reclink: an application for database linkage implementing the probabilistic record linkage method

*Kenneth R. de Camargo Jr.* [1]
*Cláudia M. Coeli* [2]

[1] *Departamento de Planejamento e Administração em Saúde, Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro. Rua São Francisco Xavier 524, 7º andar, Bloco D. Rio de Janeiro, RJ 20559-90, Brasil. kenneth@uerj.br*
[2] *Departamento de Medicina Preventiva, Faculdade de Medicina e Núcleo de Estudos de Saúde Coletiva, Universidade Federal do Rio de Janeiro. Av. Brigadeiro Trompowski s/nº, Edifício do Hospital Universitário Clementino Fraga Filho, 5º andar, Ala Sul, Rio de Janeiro, RJ 21931-059, Brasil. coeli@acd.ufrj.br*

**Abstract** *This paper presents a system for database linkage based on the probabilistic record linkage technique, developed in the C++ language with the Borland C++ Builder version 3.0 programming environment. The system was tested in the linkage of data sources of different sizes, evaluated both in terms of processing time and sensitivity for identifying true record pairs. Significantly less time was spent in record processing when the program was used, as compared to manual processing, especially in situations where larger databases were used. Manual and automatic processes had equivalent sensitivities in situations where we used databases with fewer records. However, as the number of records grew we noticed a clear reduction in the sensitivity of the manual process, but not in the automatic one. Although in its initial stage of development, the system performed well in terms of both processing speed and sensitivity. Although overall performance of algorithms was satisfactory, we intend to evaluate other routines in the attempt to improve the system's performance.*

**Key words** *Information Systems; Software; Data Comparability*

**Resumo** *Apresenta-se um sistema de relacionamento de bases de dados fundamentado na técnica de relacionamento probabilístico de registros, desenvolvido na linguagem C++ com o ambiente de programação Borland C++ Builder versão 3.0. O sistema foi testado a partir de fontes de dados de diferentes tamanhos, tendo sido avaliado em tempo de processamento e sensibilidade para a identificação de pares verdadeiros. O tempo gasto com o processamento dos registros foi menor quando se empregou o programa do que ao ser realizado manualmente, em especial, quando envolveram bases de maior tamanho. As sensibilidades do processo manual e do processo automático foram equivalentes quando utilizaram bases com menor número de registros; entretanto, à medida que as bases aumentaram, percebeu-se tendência de diminuição na sensibilidade apenas no processo manual. Ainda que em fase inicial de desenvolvimento, o sistema apresentou boa performance tanto em velocidade quanto em sensibilidade. Embora a performance dos algoritmos utilizados tenha sido satisfatória, o objetivo é avaliar outras rotinas, buscando aprimorar o desempenho do sistema.*

**Palavras-chave** *Sistemas de Informação; Software; Comparabilidade de Dados*

## Introdução

O interesse em relacionar registros em diferentes bases de dados veio aumentando nas últimas décadas juntamente com a crescente disponibilidade de grandes bases de dados em saúde informatizadas. Estas bases são empregadas muitas vezes para monitorar a ocorrência de evento de interesse (óbito, por exemplo) em estudos de coorte (Rogot et al., 1986; Van Den Brabdt et al., 1990; Horm, 1996), ou com objetivo de ampliar a quantidade de informação a ser obtida por cada unidade de estudo a partir da combinação de bases qualitativamente distintas (Newcombe et al., 1959; Dean, 1996).

Relacionar registros em diferentes bases de dados é tarefa trivial nos casos em que os registros de cada base incluam campo comum que permita a identificação de cada registro de forma unívoca, como, por exemplo, CPF. Infelizmente, um campo desta natureza raramente está presente nas bases de dados de saúde disponíveis e o processo de relacionamento deve fundamentar-se na utilização de atributos menos específicos, tais como nome, data de nascimento e endereço. A complexidade do processo cresce à medida que o número de registros a ser relacionado aumenta, tornando necessário a utilização de computadores e de rotinas automatizadas para sua execução.

Newcombe (Newcombe et al., 1959) é um dos pioneiros no desenvolvimento de metodologia para a execução do relacionamento automático de registros. Fellegi & Sunter (1969) ampliaram os conceitos originais e deram tratamento matemático formal ao método que hoje é conhecido como o método do relacionamento probabilístico de registros (Jaro, 1989; Dean, 1996). Este método se baseia em três processos, a saber: a padronização de registros, a blocagem de registros (*blocking*) e o pareamento de registros (Jaro, 1989; Dean, 1996).

O processo de padronização é o primeiro a ser realizado e envolve a preparação dos campos de dados, buscando-se minimizar a ocorrência de erros durante o processo de pareamento de registros. Este processo é fundamental para os campos não estruturados – como, por exemplo, nome – que se caracterizam pela entrada relativamente livre de controles. Exemplos de rotinas que podem ser desenvolvidas nesta etapa são a transformação de todos os caracteres alfabéticos da forma minúscula para a maiúscula, assim como a eliminação de caracteres de pontuação e de espaços em branco no início do campo. O segundo processo consiste na criação de blocos lógicos de registros dentro dos arquivos que serão relacionados. O objetivo desta etapa é permitir que o processo de pareamento se faça de forma mais otimizada. Por meio deste processo, as bases de dados são logicamente divididas em blocos mutuamente exclusivos, limitando-se as comparações aos registros pertencentes a mesmo bloco. Os blocos são constituídos de forma a aumentar a probabilidade de que os registros neles contidos representem pares verdadeiros. O processo consiste na indexação dos arquivos a serem relacionados segundo uma chave formada por um campo ou pela combinação de mais de um campo. Os registros de determinado bloco mostram idêntico valor para a chave escolhida. O bairro de residência é um tipo de campo que pode ser empregado para a blocagem de registros. Neste caso, só serão comparados os registros nas duas bases que exibem o mesmo bairro de residência.

Jaro (1989) ressalta que a escolha do campo chave deve ser feita de modo criterioso. Segundo este autor, a chave escolhida deveria permitir a divisão do arquivo no maior número de blocos possíveis – apresentar várias categorias – e, ao mesmo tempo, ser sujeita à baixa probabilidade de ocorrência de erros de registro. Por exemplo, uma blocagem a partir do campo "sexo" dividiria o arquivo em apenas dois blocos, trazendo pouco ganho em termos de otimização do processo de comparação. Já a utilização do último nome permitiria a divisão em diversos blocos, mas, como este campo é sujeito a vários erros de registro, essa estratégia de blocagem aumentaria muito a chance de que os registros relativos a mesmo indivíduo fossem classificados em blocos diferentes, impossibilitando sua comparação.

Uma das chaves empregadas usualmente nas bases de dados que trazem informações relativas ao nome é a transformação do último nome em código fonético, que é usado para a blocagem de registros. A utilização do código fonético visa minimizar o problema do erro de registro. O código fonético mais empregado é o *soundex* (Newcombe et al., 1959). Em virtude de o processo de blocagem não ser imune ao problema da classificação de registros do mesmo indivíduo em blocos diferentes, recomenda-se a utilização de estratégia de múltiplos passos (Jaro, 1989; Dean, 1996), ou seja, emprega-se determinada chave para blocagem e procede-se à comparação dos registros. Os registros não pareados na primeira etapa são então novamente blocados, empregando-se, para tanto, nova chave. Este processo pode ser repetido mais vezes, dependendo da disponibilidade de campos-chaves adequados.

O último processo envolve o pareamento de registros e baseia-se na construção de escores para os diferentes pares possíveis de serem obtidos a partir do emprego de determinada estratégia de blocagem. O conceito de escore foi inicialmente proposto por Newcombe et al. (1959) e desenvolvido posteriormente por Fellegi & Sunter (1969). Estes últimos autores propuseram, adicionalmente, a definição do conceito de escore limiar para a classificação dos pares em três categorias: verdadeiros, falsos e duvidosos. Isto significa que os pares que apresentarem o escore acima de valor predeterminado (limiar superior) serão classificados como verdadeiros, enquanto aqueles que exibirem o escore abaixo de um segundo valor também predeterminado (limiar inferior) serão considerados pares falsos. Os pares que mostrarem valores de escore intermediários entre os dois limiares serão tidos como duvidosos e deverão ser revisados manualmente.

O escore final de cada par é construído a partir da soma dos escores ponderados de cada campo empregado no processo de pareamento – por exemplo, nome, último nome, sexo e data de nascimento –, permitindo, desta maneira, que cada campo contribua de modo diferenciado para o escore total do par. A contribuição diferenciada é recomendável, pois os campos apresentam diferente poder discriminatório e, ao mesmo tempo, exibem maior ou menor probabilidade de terem seus conteúdos registrados de forma incorreta. Por exemplo, o campo sexo mostra baixo poder discriminatório, mas o seu registro é, em geral, feito de forma correta. Já o campo "último nome", apesar de apresentar bom poder discriminatório, é mais sujeito a erros de registro.

Os pesos são construídos a partir de conceitos muito utilizados entre os epidemiologistas na avaliação da acurácia de testes diagnósticos. Para cada campo $i$ define-se a probabilidade $m_i$ do campo concordar entre os dois registros, dado que se trata de par verdadeiro, e a probabilidade $u_i$ do campo concordar por trata-se de par falso. Em outras palavras, $m_i$ representa a probabilidade do campo identificar um par como verdadeiro quando ele realmente é verdadeiro (sensibilidade) e $u_i$ a probabilidade do campo identificar um par como verdadeiro, quando na realidade ele é falso (1 – especificidade). De forma análoga, poder-se-ia definir $1 - m_i$ como a probabilidade de o campo discordar entre dois registros, uma vez que se trata de par verdadeiro (1 – sensibilidade), enquanto que $1 - u_i$ representaria a probabilidade de o campo discordar, já que se trata de par falso (especificidade).

Com base nestas probabilidades são construídos dois fatores de ponderação: um, para a situação de concordância e outro, para a situação de discordância. Ou seja, compara-se o campo do primeiro registro com o do segundo registro. Se os campos concordarem, aplica-se o fator de ponderação de concordância e, em caso contrário, o de discordância. O fator de ponderação de concordância é calculado como o logaritmo de base 2 da razão de verossimilhança entre as probabilidades $m_i$ e $u_i$ ($wc_i = \log_2 \left[ \frac{m_i}{u_i} \right]$)

e o de discordância como o logaritmo de base 2 da razão de verossimilhança entre as probabilidades $1 - m_i$ e $1 - u_i$ ($wd_i = \log_2 \left[ \frac{(1 - m_i)}{(1 - u_i)} \right]$)

O escore total de determinado par é obtido a partir da soma dos fatores de ponderação atribuídos após a comparação de cada campo avaliado. Como $m_i$ é geralmente maior que $u_i$, o fator de concordância contribui positivamente para o escore final, enquanto o fator de discordância contribui negativamente (Jaro, 1989).

Jaro (1989) chama a atenção para o fato de nem sempre ser fácil a decisão acerca da concordância ou discordância entre dois campos de determinado par. Conseqüentemente, muitas vezes é difícil escolher qual o fator de ponderação a ser atribuído como resultado da comparação de dois campos. Por exemplo, considere-se que o primeiro registro de um par em avaliação apresentasse a data de nascimento igual a "03/07/29", enquanto no segundo, esta data fosse igual a "29/01/35". Estas datas são bem diferentes e não haveria problema em atribuir o fator de ponderação de discordância. Entretanto, que decisão deveria ser tomada no caso de a data do segundo registro ser igual a "06/07/29"? É verdade que estas datas não são exatamente iguais, mas, sem sombra de dúvida, poder-se-ia considerar que "concordam" razoavelmente bem. Jaro (1989) propõe, como solução para os casos em que a discordância é pequena, a atribuição do fator de ponderação de concordância, porém não de forma integral. Ou seja, atribui-se um valor que contribuirá positivamente para o escore final, porém esta contribuição será menor do que aquela que seria utilizada no caso de a concordância ser exata. Resta, portanto, definir qual a discordância "aceitável" e que parcela do valor do fator de ponderação de concordância que deve ser empregada. Estas definições dependerão do tipo do campo avaliado e do algoritmo de comparação utilizado e certamente constituem interessante área de investigação a ser explorada.

Por fim, os valores de $m_i$ e $u_i$, assim como os valores de limiar superior e inferior podem ser estimados. Fellegi & Sunter (1969) e Jaro (1989) discutem uma metodologia para a estimativa destes parâmetros. De forma mais simples, também podem ser empregados valores previamente conhecidos pelo pesquisador. Dean (1996) considera que podem ser utilizados valores em torno de 0,9 e 0,1, para $m_i$ e $u_i$, respectivamente, para a maioria dos campos. É claro que existem exceções, como é o caso do campo "sexo", no qual seria melhor empregar $u_i = 0{,}5$. Este campo admite apenas dois valores, havendo, portanto, 50% de probabilidade de o mesmo concordar apenas em função do acaso.

Neste trabalho buscou-se avaliar um sistema de relacionamento de bases de dados fundado na técnica de relacionamento probabilístico de registros (*probabilistic record linkage*), desenvolvido na linguagem C++ com o ambiente de programação *Borland* C++ *Builder* versão 3.0.

## Metodologia

O *software*, denominado *RecLink*, foi desenvolvido na linguagem C++ com o ambiente de programação *Borland* C++ *Builder* versão 3.0 (Borland International Inc., 1998a; Reisdorph, 1998). O programa consiste em uma interface com bancos de dados flexível que permite ao usuário designar, de modo interativo, as regras de associação entre duas tabelas. O processo opera em dois níveis: primeiramente criam-se blocos de registros (*blocking)*, com base no código soundex dos campos selecionados (em princípio, contendo nomes) e, dentre os registros blocados segundo mesmo código, outras variáveis (denominadas pareamento, podendo variar de uma a três) são usadas para atribuir peso numérico à associação dos registros. Na atribuição de pesos, três algoritmos diferentes podem ser utilizados na comparação dos respectivos campos: a comparação pura e simples, que só retorna valor verdadeiro caso o conteúdo seja rigorosamente idêntico, a comparação de seqüências de caracteres caractere a caractere e a comparação aproximada. A Figura 1 mostra a tela de configuração das opções do sistema.

O programa foi avaliado a partir dos dados coletados por um dos autores (Coeli, 1998) para a realização de estudo que visava à avaliação da factibilidade para a implantação de sistema de vigilância do diabetes mellitus na população idosa residente na Área Programática 2.2 da cidade do Rio de Janeiro. As fontes de dados empregadas para este estudo foram: Sistema sobre Mortalidade (SIM), Sistema de Informações Hospitalares do SUS (SIH-SUS) e estatísticas ambulatoriais provenientes de unidades de saúde tidas como sentinelas para a captação de casos de diabetes mellitus na área.

As bases foram padronizadas com a transformação da data de nascimento em um campo caractere com seis posições, e com a utilização de idêntica regra de codificação do campo sexo nos diferentes arquivos. O campo nome foi processado visando à transformação de todos os caracteres alfabéticos da forma minúscula para a maiúscula, a eliminação de caracteres de pontuação e a eliminação de espaços em branco no início do campo. A padronização dos campos data de nascimento e sexo foi realizada mediante o emprego do gerenciador de banco de dados *Visual dBASE* (Borland International Inc., 1998b), ao passo que, para o nome, utilizou-se uma rotina informatizada especialmente desenvolvida para este fim (Camargo Jr., 1997).

Duas estratégias foram empregadas para o relacionamento de registros. Na primeira foram utilizados os arquivos ambulatorial (363 registros), hospitalar (134 registros) e de mortalidade (291 registros), relativos aos registros dos pacientes com diagnóstico de diabetes. Na segunda foram empregados o arquivo ambulatorial acima citado (363 registros), o arquivo com os óbitos por diabetes relativos ao ano de 1994 (190 registros), o arquivo completo dos óbitos em idosos residentes na AP 2.2 (ano 95 – 2.857 registros) e o arquivo completo das hospitalizações de idosos residentes na Cidade do Rio de Janeiro (anos de 1994 e 1995 – 79.039 registros).

Inicialmente realizou-se manualmente o processo de relacionamento. Ou seja, os arquivos utilizados em cada estratégia foram combinados, resultando em arquivo com 740 registros na primeira estratégia e em arquivo com 82.449 registros na segunda. Estes arquivos foram indexados por nome e data de nascimento, tendo sido efetuada a inspeção visual para a identificação de registros pertencentes a mesmo paciente. Quando um par era identificado, os registros completos eram revisados para que outros dados permitissem a confirmação de que se tratava de par verdadeiro. Os dados empregados para este fim foram: a data do óbito; a data da alta e o tipo da alta hospitalar; a data do último atendimento ambulatorial; e, por fim, o endereço de residência.

A seguir empregou-se o programa *RecLink*, utilizando-se três estratégias consecutivas de

blocagem. Em primeiro lugar foi feita a blocagem pela combinação dos códigos *soundex* do último e do primeiro nome. No passo seguinte foi realizada a blocagem pelo *soundex* do primeiro nome e por fim, usou-se a blocagem pelo *soundex* do último nome. Não se empregou procedimento formal para a estimativa dos valores dos parâmetros (Fellegi & Sunter, 1969; Jaro 1989). Estes foram selecionados após alguns testes com subconjuntos da base de dados utilizada para este estudo, partindo dos valores sugeridos pela bibliografia (Dean, 1996). Tais valores são apresentados na Tabela 1.

Em cada passo foram calculados o tempo de processamento para a execução das comparações, o total de pares no bloco, o total de pares com escore positivo e, dentre estes, a proporção de pares verdadeiramente positivos (valor preditivo positivo no bloco). Quanto ao último item, todos os pares com escore positivo que não haviam sido identificados durante a busca manual foram revisados para confirmação da natureza do par (verdadeiro ou falso positivo). A sensibilidade das três etapas de blocagem em conjunto foi estimada, adotando-se, como padrão ouro, o total de pares verdadeiros identificados tanto durante a busca manual como durante a automática, e calculando-se adicionalmente os respectivos intervalos de confiança exatos de 90% (Stata Corporation, 1997).

## Resultados

Na Tabela 2 são expostos os resultados obtidos com a realização do relacionamento das bases de dados por intermédio da utilização do programa *RecLink*. Os três primeiros relacionamentos envolveram apenas as bases de dados cujos registros referiam-se a pacientes com o diagnóstico de diabetes registrado (primeira estratégia de relacionamento), enquanto foram empregadas nos demais ou a base de internações completa de idosos residentes no Município do Rio de Janeiro (anos de 1994 e 1995), ou a base completa de mortalidade do ano de 1995 relativa aos idosos residentes na AP 2.2 associada à base de mortalidade (ano 1994) por diabetes de idosos residentes na AP 2.2 (segunda estratégia de relacionamento).

Para cada relacionamento efetuado são apresentados o número de pares possíveis de serem obtidos em função do número de registros existentes em cada base de dados envolvida no relacionamento, o número de pares efetivamente avaliados em cada estratégia de blocagem empregada, o tempo consumido com o processamento, o número de pares com escore positivo no bloco, e dentre estes, a proporção representada pelos pares verdadeiros (valor preditivo positivo no bloco). Nos passos 2 e 3 somente foram avaliados os registros considerados não "pareados" nas etapa anteriores, ou

Figura 1

Tela de configuração do programa *RecLink*. À esquerda vêem-se as opções de blocagem e, à direita, as opções de pareamento para uma das variáveis (ver texto).

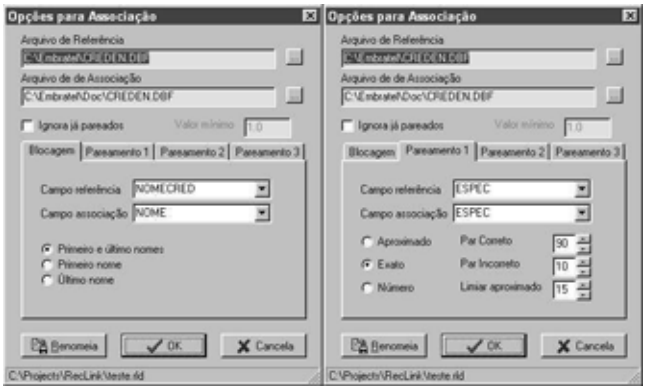


Tabela 1

Parâmetros utilizados no teste do programa *RecLink*.

| Campo | Algoritmo | Sensibilidade $m_i$ | 1- especificidade $u_i$ | Proporção mínima de concordância |
|---|---|---|---|---|
| Nome | Aproximado | 90% | 5% | 85% |
| Data de nascimento | Número | 90% | 10% | 65% |
| Sexo | Exato | 95% | 50% | — |

Tabela 2

Resultados obtidos com a realização do relacionamento de registros utilizando o programa *RecLink*.

| Arquivos relacionados | | Número de pares possíveis | Blocagem* | Tempo de processamento | Total de pares no bloco | Pares no bloco com Escore positivo | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Total | Verdadeiros positivos | |
| | | | | | | | n | % |
| **Hospitalizações** | **Mortalidade** | 38.994 | Passo 1 | < 30 s | 46 | 30 | 29 | 96,7 |
| **por diabetes** | **por diabetes** | | Passo 2 | < 30 s | 547 | 2 | 2 | 100 |
| (134 registros) | (291 registros) | | Passo 3 | < 30 s | 562 | 0 | — | — |
| Total de pares verdadeiros não encontrados*** | | | 2 | | | | | |
| **Hospitalizações** | **Ambulatório/** | 48.642 | Passo 1 | < 30 s | 39 | 11 | 11 | 100 |
| **por diabetes** | **Diabetes** | | Passo 2 | < 30 s | 1.242 | 1 | 0 | 0 |
| (134 registros) | (363 registros) | | Passo 3 | < 30 s | 921 | 0 | — | — |
| Total de pares verdadeiros não encontrados*** | | | 1 | | | | | |
| **Mortalidade** | **Ambulatório/** | 105.633 | Passo 1 | < 30 s | 50 | 6 | 5 | 83,3 |
| **por diabetes** | **Diabetes** | | Passo 2 | < 30 s | 2.527 | 1 | 0 | 0,0 |
| (291 registros) | (363 registros) | | Passo 3 | < 30 s | 1.673 | 3 | 1 | 33,3 |
| Total de pares verdadeiros não encontrados*** | | | 0 | | | | | |
| **Hospitalizações** | **Mortalidade por** | 408.298 | Passo 1 | < 1 min | 162 | 29 | 29 | 100 |
| **por diabetes** | **todas as causas** | | Passo 2 | < 1min | 4.729 | 8 | 1 | 12,5 |
| (134 registros) | (3.047 registros)** | | Passo 3 | < 1 min | 4.726 | 10 | 1 | 10,0 |
| Total de pares verdadeiros não encontrados*** | | | 3 | | | | | |
| **Ambulatório/** | **Mortalidade por** | 1.106.061 | Passo 1 | < 1 min | 381 | 13 | 9 | 69,2 |
| **Diabetes** | **todas as causas** | | Passo 2 | < 1min | 21.550 | 37 | 1 | 2,7 |
| (363 registros) | (3.047 registros)** | | Passo 3 | < 1 min | 13.237 | 35 | 2 | 5,7 |
| Total de pares verdadeiros não encontrados*** | | | 0 | | | | | |
| **Mortalidade** | **Hospitalizações/** | 23.000.349 | Passo 1 | < 1 min | 8.581 | 222 | 107 | 48,2 |
| **por diabetes** | **todas as causas** | | Passo 2 | 7 min | 191.808 | 465 | 3 | 0,6 |
| (291 registros) | (79.039 registros)** | | Passo 3 | 5 min | 144.582 | 209 | 2 | 1,0 |
| Total de pares verdadeiros não encontrados*** | | | 4 | | | | | |
| **Ambulatório/** | **Hospitalizações/** | 28.691.157 | Passo 1 | < 1 min | 18.251 | 359 | 45 | 12,5 |
| **Diabetes** | **todas as causas** | | Passo 2 | 10 min | 325.345 | 1.117 | 0 | 0 |
| (363 registros) | (79.039 registros)** | | Passo 3 | 8 min | 212.233 | 701 | 2 | 0,3 |
| Total de pares verdadeiros não encontrados*** | | | 0 | | | | | |

* Passo 1: combinação do *soundex* do último e do primeiro nome; Passo 2: *soundex* do primeiro nome; Passo 3: *soundex* do último nome.
** Bases completas: mortalidade em idosos residentes da AP 2.2 (1995) + Mortalidade por diabetes em idosos residentes na AP 2.2 (1994); Hospitalizações de idosos residentes na Cidade do Rio de Janeiro (1994 e 1995).
*** Registros identificados através da busca manual prévia.

seja, aqueles que não foram incluídos nos blocos anteriores ou que mostraram escore negativo. Por fim, também é apresentado o número de pares verdadeiros, os quais haviam sido identificados previamente bela busca manual, mas que não foram identificados pelo programa após a execução dos três passos de blocagem.

Em todos os relacionamentos realizados observa-se importante redução do número de pares efetivamente avaliados dentro dos blocos, o que contribuiu para a diminuição do tempo total necessário ao processamento das comparações. Mesmo naqueles relacionamentos que envolveram bases grandes, o tempo para o processamento dos blocos foi de, no máximo, dez minutos. Também pode ser observada tendência de diminuição acentuada do valor preditivo positivo dentro dos blocos nos relacionamentos que envolvem as maiores bases de dados.

Na Tabela 3 são apresentadas as sensibilidades para a identificação de pares verdadeiros alcançadas por meio da busca manual e da utilização do programa *RecLink*. Nesta análise foi considerado, como padrão ouro, o total de pares verdadeiros identificados nos dois processos em conjunto. Verifica-se que a performance do programa foi equivalente à da busca manual quando foram usadas as bases com pequeno número de registros (primeira estratégia de relacionamento). Entretanto, o programa alcançou resultados melhores para os relacionamentos que envolvem as bases com maior número de registros (segunda estratégia de relacionamento).

## Discussão

Os resultados obtidos com o *software* foram bastante razoáveis, principalmente ao se considerar que este se encontra na versão inicial e que não se empregou metodologia mais formal para a estimativa dos valores dos parâmetros de trabalho. O tempo consumido com o processamento dos registros foi expressivamente menor ao ser utilizado o programa do que quando se realizou o processamento manual, em especial, nos relacionamentos que envolveram as bases de maior tamanho.

As sensibilidades do processo manual e do processo automático foram equivalentes nas situações em que foram usadas as bases com menor número de registros; entretanto, à medida que as bases relacionadas aumentaram,

Tabela 3

Sensibilidade para a identificação de pares verdadeiros. Comparação entre a busca manual e a utilização do programa *RecLink*.

| **Arquivos relacionados** | | **Número de pares verdadeiros** | **Busca manual** Sensibilidade | I.C. 90% | **Programa *RecLink*** Sensibilidade | I.C. 90% |
|---|---|---|---|---|---|---|
| Hospitalizações por diabetes | Mortalidade por diabetes | 33 | 100% | 89,8%-100% | 93,9% | 81,2%-98,7% |
| Hospitalizações por diabetes | Ambulatório/ Diabetes | 12 | 100% | 75,4%-100% | 91,7% | 64,8%-99,4% |
| Mortalidade por diabetes | Ambulatório/ Diabetes | 6 | 100% | 58,6%-100% | 100% | 58,6%-100% |
| Hospitalizações por diabetes | Mortalidade por todas as causas | 34 | 90,9% | 74,5%-97,6% | 90,9% | 74,5%-97,6% |
| Ambulatório/ Diabetes | Mortalidade por todas as causas | 12 | 50,0% | 25,2%-74,8% | 100% | 75,4%-100% |
| Mortalidade por Diabetes | Hospitalizações por todas as causas | 115 | 70,4% | 62,6%-77,3% | 96,5% | 91,9%-98,7% |
| Ambulatório/ Diabetes | Hospitalizações por todas as causas | 47 | 66,0% | 52,9%-77,1% | 100% | 92,7%-100% |

percebeu-se tendência de diminuição na sensibilidade do processo manual, mas não no processo automático. O processo manual foi desenvolvido a partir da combinação dos diferentes arquivos em arquivo único, que depois foi indexado por nome e data de nascimento. Após o aumento de tamanho dos arquivos, também cresceu o número de registros com nomes parecidos. Por exemplo, se um indivíduo chamado "Manoel" tivesse seu nome digitado em uma base com "o" e, na outra, com "u" ("Manoel" x "Manuel"), o número de registros que separaria os dois nomes impediria ou, ao menos, dificultaria bastante identificá-los como pertencentes a mesma pessoa.

A metodologia proposta por Fellegi & Sunter (1969) orienta a definição de um escore limiar superior e outro inferior para a classificação dos pares. Pares com escore acima do limiar superior seriam classificados como verdadeiros, enquanto aqueles com escore inferior seriam classificados como falsos. Já os com escore intermediário seriam classificados como duvidosos, devendo ser revisados. No presente trabalho definiu-se valor arbitrário apenas para o limiar inferior, ou seja, considerou-se que os registros com escore negativo eram falsos e que os registros com escore positivo deveriam ser revistos. A definição adicional de um limiar superior, entretanto, é fundamental para otimizar o processo do relacionamento que envolve bases de maior tamanho. Este fato torna-se evidente ao observar-se o crescimento do número de pares com escore positivo e a diminuição do valor preditivo positivo nos relacionamentos que envolveram as bases de maior tamanho. Por outro lado, a estimativa dos valores de limiar que emprega procedimentos mais formais, permitiria definir o conjunto de valores que viria a alcançar o melhor equilíbrio entre a otimização do processo e a acurácia dos resultados.

A diminuição do valor preditivo positivo para os relacionamentos que abrangem as bases com maior tamanho é possivelmente determinada por dois fatores: a diminuição da prevalência dos pares verdadeiros e a diminuição da especificidade global do processo de comparação. Nos relacionamentos que envolvendo as bases menores, todos os registros referiam-se a pacientes que apresentavam diagnóstico de diabetes registrado. Em outras palavras, a seleção segundo diagnóstico fez com que os registros das bases menores sofressem blocagem prévia, acrescendo a prevalência da ocorrência de pares verdadeiros. Por sua vez, é provável que o aumento do número de registros também tenha modificado o "perfil" de nomes e das datas de nascimento, isto é, poder-se-ia esperar o aumento da freqüência de pacientes com nomes e datas de nascimento semelhantes, o que, por sua vez, poderia levar à diminuição da especificidade das comparações. É interessante ressaltar que a sensibilidade não foi modificada, pois os resultados falsos negativos ocorreriam em decorrência de erros na hora de informar ou registrar os dados (o nome, por exemplo), sendo que as freqüências destes erros tenderiam, em princípio, a se manter constantes.

Estes últimos fenômenos são afins aos observados quando se avalia um teste diagnóstico em uma população na qual 50% dos pacientes são doentes e 50% são normais. Se este teste for posteriormente aplicado a uma população em que, ao invés de 50% de pessoas normais, existirem pessoas com outras doenças que possam exibir manifestações semelhantes às da doença em questão, então pode ser que a especificidade do teste nesta nova população venha a se modificar. De forma análoga, a sensibilidade também poderia se modificar se a nova população apresentasse proporção maior de pacientes com formas leves da doença (Sackett et al., 1985).

Quanto melhor o teste, mais robusto ele será frente a estes tipos de mudanças. Em termos do programa proposto, esta robustez poderia ser alcançada mediante a utilização de algoritmos de comparação mais eficientes. Jaro (1989) desenvolveu um algoritmo para a comparação de campos caractere que leva em conta a inserção, deleção, troca e transposição aleatória de caracteres. Infelizmente não se teve acesso ao mesmo, pois é parte de um *software* comercial para o relacionamento probabilístico de registros (AutoStan-AutoMatch; MatchWare Technologies, Inc.), cuja licença para versão mais barata custa US$ 12,500.

No programa *RecLink* empregou-se um algoritmo de busca aproximada (*fuzzy search*) para a comparação de campos caracteres, ao passo que para os campos "data" utilizou-se um algoritmo mais simples, que faz comparações de seqüências de dígitos. O algoritmo de busca aproximada utilizado nesta primeira versão é adaptação de uma rotina originalmente destinada à busca de seqüências de caracteres em textos, disponível para *download* gratuito em http://www.snippets.org. A versão original retornava dois parâmetros: um ponteiro para localização no texto da seqüência procurada e um valor numérico que expressava o ajuste do texto localizado com a seqüência buscada, retornando 0 a um ajuste perfeito e um inteiro igual ao número de caracteres da seqüência de busca, caso esta não tivesse sido encontrada. A adaptação descartou o primeiro parâmetro e

traduziu o segundo para um número de ponto flutuante, que varia entre 0 (desajuste total) e 1 (ajuste perfeito).

Embora a performance dos algoritmos em conjunto tenha sido boa, é objetivo deste estudo avaliar outras alternativas, como, por exemplo, trabalhar com variações da idade ao invés de comparar os dígitos da data. Durante as simulações realizadas, percebeu-se que o algoritmo de busca aproximada mostra queda de desempenho nas comparações com nomes que incluam preposições, por exemplo, "João da Silva". Assim sendo, uma rotina para a exclusão destas preposições será incluída nas próximas versões do programa. Adicionalmente, para dado conjunto de algoritmos, melhor performance do programa seria alcançada a partir da estimativa dos valores dos parâmetros de trabalho, utilizando-se, para tal, a metodologia proposta por Fellegi & Sunter (1969).

## Referências


BORLAND INTERNATIONAL INC., 1998a. *Borland C++ Builder 3 Developer's Guide*. Scotts Valley: Borland International Inc.

BORLAND INTERNATIONAL INC., 1998b. *Visual dBase Version 5.6*. CD-ROM. Scotts Valley: Borland International Inc.

CAMARGO Jr., K. R., 1997. *DBFCHECK – Revisor de Campos Alfabéticos*. Rio de Janeiro: Instituto de Medicina Social, Universidade Estadual do Rio de Janeiro. (mimeo.)

COELI, C. M., 1998. *Vigilância do Diabetes Mellitus em uma População Idosa: Aplicação da Metodologia de Captura-Recaptura*. Tese de doutorado, Rio de Janeiro: Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro.

DEAN, J. M., 1996. Probabilistic Linkage of Records. 15 november 1997 <http://www.nedarc.med.utah.edu/nedarc/linkage/description/prob-links/prob2.htlm>.

FELLEGI, I. P. & SUNTER, A. B., 1969. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183-1210.

HORM, J., 1996. Linkage of the National Health Interview Survey with the National Death Index: Methodologic and Analytic Issues. 15 november 1997 <htpp://www.cpc.unc.edu/pubs/paa_papers/1996/horm.htlm>.

JARO, M. A., 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84:414-420.

NEWCOMBE, H. B.; KENNEDY, J. M.; AXFORD, S. J. & JAMES, A. P., 1959. Automatic linkage of vital records. *Science*, 130:954-959.

REISDORPH, K., 1998. *Teach yourself Borland C++ Builder 3 in 14 days*. Indianapolis: SAMS Publishing.

ROGOT, E.; SORLIE, P. & JOHNSON, N. J., 1986 Probabilistic methods in matching census samples to the National Death Index. *Journal of Chronic Diseases*, 39:719-734.

SACKETT, D.; HAYNES, B. & TUGWELL, P., 1985. *Clinical Epidemiology*. Boston: Little Brown & Co.

STATA CORPORATION, 1997. *Stata Reference Manual: Release 5, Volumes 1-3*. College Station: Stata Press.

VAN DEN BRABDT, P. A.; SCHOUTEN, L.; GOLDBOHM, R. A.; DORANT, E. & HUNEN, P. M. H., 1990. Development of a record linkage protocol for use in the Dutch Cancer Registry for epidemiological research. *International Journal of Epidemiology*, 19:553-558.