

# cadernos Saúde Coletiva

## Relacionamento de Bases de Dados em Saúde

EDITORES CONVIDADOS

Claudia Medina Coeli

Kenneth Rochel de Camargo Jr.

Catálogo na fonte – Biblioteca do CCS / UFRJ

---

Cadernos Saúde Coletiva / Universidade Federal do Rio de Janeiro,  
Núcleo de Estudos de Saúde Coletiva, v.XIV, n.2 (abr . jun 2006).

Rio de Janeiro: UFRJ/NESC, 1987-.

Trimestral

ISSN 1414-462X

1.Saúde Pública - Periódicos. I I.Núcleo de Estudos de Saúde Coletiva/UFRJ.

---

## EDITORIAL

### RELACIONAMENTO DE BASES DE DADOS EM SAÚDE

Rejane Sobrino Pinheiro<sup>1</sup>, Kenneth Rochel de Camargo Jr.<sup>2</sup>,  
Claudia Medina Coeli<sup>3</sup>

Susser e Susser (1996) chamaram a atenção sobre a importância para a pesquisa epidemiológica de duas áreas recentes de desenvolvimento tecnológico: as técnicas em biologia e biomedicina, que vêm contribuindo para a melhor compreensão dos processos de determinação de doenças no micronível; e as tecnologias na área de sistemas de informação, que ao permitirem o armazenamento e acesso em larga escala (geográfica e no tempo) a bases de dados com informações variadas sobre condições de vida e saúde, têm possibilitado o melhor entendimento dos processos de determinação de doenças no macronível.

O Brasil possui grandes bases de dados nacionais, de dados vitais, de morbidade e de produção de serviços, de abrangência nacional, comparáveis às que existem em diversos países centrais. É produzido anualmente um grande volume de dados, amplamente disponíveis via internet pelo Datasus/MS, cujo percentual de aproveitamento deixa ainda a desejar. Parte da argumentação gira em torno da qualidade dos dados, outra parte é que os dados não são detalhados o suficiente para serem úteis para apoiar a decisão em saúde, e há, ainda, a falta de treinamento por parte dos gestores no uso da informação. Todo sistema de informação tem uma curva de implantação e, a partir do uso dos dados processados e do retorno da informação aos diferentes níveis de utilização, chegando até os que a coletam, se dá o constante processo de melhoria dos dados e do próprio sistema de informações.

Lazaridis (1998) aponta que o uso de bases de dados na avaliação em saúde segue o mantra dos três Rs – “Redução de custos; Reutilização e Reciclagem”.

<sup>1</sup> Doutora em Saúde Pública. Prof<sup>a</sup>. Adjunta do Núcleo de Estudos de Saúde Coletiva e do Departamento de Medicina Preventiva da Faculdade de Medicina - UFRJ - End.: Av. Brigadeiro Trompowsky, s/nº - Ilha do Fundão - Praça da Prefeitura da Cidade Universitária - CEP: 21949-900 - Rio de Janeiro - RJ e-mail: rejane@nesc.ufrj.br

<sup>2</sup> Doutor em Saúde Coletiva. Prof. do Instituto de Medicina Social - IMS/UERJ - End.: Rua São Francisco Xavier 524, 7º andar Bl. D - CEP: 20559-900, Rio de Janeiro, RJ - e-mail: kenneth@uerj.br

<sup>3</sup> Doutora em Saúde Coletiva. Prof<sup>a</sup>. Visitante do Departamento de Epidemiologia. Instituto de Medicina Social - IMS/UERJ e Pesquisadora Visitante da Escola Politécnica de Saúde Joaquim Venâncio/FIOCRUZ. e-mail: coelicm@terra.com.br

Segundo o autor, essas fontes de dados permitem que avaliações em saúde sejam realizadas a um custo reduzido, já que dados colhidos anteriormente são reutilizados para a condução de análises que não haviam sido originalmente planejadas (a idéia de reciclagem). Aos três Rs iniciais, o autor propõe a inclusão de um quarto R, para representar a necessidade do uso responsável dessas fontes no que diz respeito aos aspectos éticos e legais (particularmente em termos de integridade e segurança dos dados). Outro sentido, entretanto, que poderia ser dado para a idéia da responsabilidade, envolve o cuidado que pesquisadores devem ter ao usarem fontes de dados secundários. Pesquisas baseadas na coleta de dados primários partem de uma pergunta específica a ser respondida, sendo, então, utilizados um conjunto de procedimentos para que todos os dados necessários à análise sejam coletados e armazenados de forma adequada. Já na pesquisa baseada em dados secundários, a questão que se coloca é buscar que perguntas podem ser respondidas considerando a qualidade e a natureza dos dados disponíveis na base de dados selecionada para a análise. As técnicas de relacionamento de registros, ao permitirem a integração de bases de dados de natureza diversa, ampliam o escopo de perguntas a serem respondidas, além de contribuir para a melhoria da qualidade dos dados registrados e permitir o seguimento longitudinal.

Este número temático da Cadernos Saúde Coletiva busca ampliar o debate em torno da integração de bases de dados em saúde, reunindo trabalhos acerca das diversas análises que podem ser feitas, tocando em temas como a avaliação da qualidade dos dados, complementação e correção dos mesmos, como em relação à pesquisa em serviços de saúde. Aborda ainda a complexidade metodológica, tanto no desenvolvimento quanto no uso adequado das metodologias disponíveis para a integração de bases de dados, como no volume de trabalho manual que ainda é necessário para efetuar essa integração.

Esperamos que este número auxilie a consolidação e expansão deste recurso técnico-metodológico em nosso meio, pelo potencial que apresenta tanto em termos de pesquisa quanto para a avaliação e gestão de serviços de saúde.

## REFERÊNCIAS BIBLIOGRÁFICAS

LAZARIDIS, E. M. Database standardization, linkage, and the protection of privacy. *Annals Internal Medicine*. v. 127(8 Pt 2), p. 696, 1997.

SUSSER, M.; SUSSER, E. Choosing a future for epidemiology: I. Eras and paradigms. *American Journal of Public Health*. v. 86, n. 5, p. 668 - 673, 1996.

## NOTAS

### **RecLink 3: NOVA VERSÃO DO PROGRAMA QUE IMPLEMENTA A TÉCNICA DE ASSOCIAÇÃO PROBABILÍSTICA DE REGISTROS (*PROBABILISTIC RECORD LINKAGE*)**

*RecLink 3: a new version of the program that implements the probabilistic record linkage technique*

Kenneth Rochel de Camargo Jr<sup>1</sup>, Claudia Medina Coeli<sup>2</sup>

#### RESUMO

Este artigo apresenta as principais modificações e inovações da versão 3 do programa **RecLink**, que implementa a técnica de associação probabilística de registros. As principais áreas abordadas nesta revisão foram a melhora na usabilidade; melhora na rotina de combinação e implementação de rotina de verificação de duplicidades. O artigo apresenta cada uma destas em detalhes.

#### PALAVRAS-CHAVE

Relacionamento probabilístico, banco de dados, *software*

#### ABSTRACT

This paper presents the main modifications and improvements for the third version of the **RecLink** software, a program that implements the probabilistic record linkage technique. The main areas dealt with in this revision were improvements in usability; improvements in the joining routine and implementation of a routine for checking duplicities. The paper describes each of these in detail.

#### KEY WORDS

Probabilistic record linkage, database, software

#### INTRODUÇÃO

O programa RecLink (Camargo & Coeli, 2000), que implementa a técnica de associação probabilística de registros (Fellig & Sunter, 1969) teve o início do seu desenvolvimento em 1998. Desde a primeira versão, sua base de usuários progressivamente se expandiu para além dos círculos acadêmicos em direção às aplicações na área de gestão em saúde, em especial na vigilância e avaliação em saúde. Com os novos usos do programa com grandes bases de dados surgiram novas necessidades, que se constituíram no principal motor para esta revisão.

<sup>1</sup> Doutor em Saúde Coletiva. Professor Adjunto do Instituto de Medicina Social - (IMS/UERJ). End.: Rua São Francisco Xavier 524, 7º andar Bl. D - CEP: 20550-900, Rio de Janeiro, RJ - e-mail: kenneth@uerj.br

<sup>2</sup> Doutora em Saúde Coletiva. Profª. Visitante do Departamento de Epidemiologia. Instituto de Medicina Social – IMS/UERJ e Pesquisadora Visitante da Escola Politécnica de Saúde Joaquim Venâncio/FIOCRUZ.

O objetivo deste artigo é apresentar a nova versão do programa, com uma descrição dos principais acréscimos e modificações.

### O PROGRAMA REC LINK III

A tela de abertura do programa é apresentada na Figura 1. Três aspectos foram fundamentais no desenho da nova versão: melhora na usabilidade (isto é, facilitar o uso do programa); melhora na rotina de combinação; implementação de rotina de verificação de duplicidades (ver mais adiante). Como nas versões anteriores, esta foi desenvolvida na linguagem C++, utilizando, nesta versão, o ambiente de programação *Borland Development Studio* (2006).



Figura 1  
Tela inicial do programa.

### ASSISTENTES

Visando obter uma interface mais amigável para o usuário, a nova versão do programa sofreu uma modificação radical nas telas relativas às diferentes rotinas, que passaram a contar com **assistentes**.

Assistentes são usados com frequência em programas para Windows®; são uma sequência de telas que orientam passo a passo a execução de uma dada rotina. No caso do RecLink III, as várias funções foram separadas em **configuração** e **execução**; a primeira é feita por assistentes que orientam o usuário na seleção dos vários parâmetros necessários à operação da rotina, e informam ao final do processo se há alguma omissão ou erro no conjunto de opções selecionadas. Caso não haja nenhum problema, é possível salvar a configuração produzida; caso contrário, o assistente informa quais são os erros detectados. Os arquivos de configuração são arquivos de texto, com extensões diferenciadas de acordo com a rotina. Um exemplo de assistente para a rotina de identificação de duplicidades é apresentado na Figura 2.



Figura 2

Assistente da rotina para eliminação de duplicidades.

As opções de execução das rotinas solicitam a informação do arquivo de configuração previamente selecionado e eventuais parâmetros adicionais necessários, e informam ao usuário sobre o andamento da operação.

### VISUALIZAÇÃO

A rotina para a visualização de arquivos sofreu modificação importante na nova versão do programa, passando a permitir a inspeção de arquivos ordenados segundo diferentes índices (Figura 3). Ainda nesta tela é possível agora acrescentar índices ao arquivo em uso (clcando-se no botão *Novo*, no canto superior direito), o que facilita sua inspeção tanto nesta rotina quanto na combinação de arquivos.



Figura 3

Tela da rotina que permite a visualização de arquivos.

## COMBINAÇÃO

Na nova versão do programa, as rotinas para combinação de arquivos, seleção de pares verdadeiros e geração de arquivos reduzidos para os passos de blocagem subsequentes foram combinadas em uma só. Além da economia de tempo para a realização desses processos, a rotina apresenta lado a lado os registros a serem inspecionados, o que facilita o processo de revisão manual e seleção de pares verdadeiros (Figura 4).



Figura 4  
Tela da rotina de combinação e seleção de pares verdadeiros.

Esta rotina baseia-se na configuração criada pelo assistente de relacionamento para identificar os arquivos de referência e comparação, e utiliza o arquivo gerado pela rotina de relacionamento para apresentar os escores alcançados. Este último arquivo, que já existia na versão anterior, passou a incluir um campo que identifica seu atributo para combinação como par, dúvida, não par ou ignorado (a opção *default*). A nova rotina de combinação, a partir da marcação (automática e/ou manual) dos pares identificados na tela da Figura 4 gera automaticamente três arquivos: um combinado dos dois arquivos pareados e os arquivos derivados desses, que excluem os registros já pareados, que serão utilizados no passo seguinte, para o novo pareamento. Nos testes em andamento este novo desenho da rotina permitiu uma economia considerável de tempo nas operações de relacionamento em passos sucessivos.

## CÁLCULO DE ESCORES

Diversas rotinas do RecLink incluem como parâmetro um escore, seja para visualização, seja para inclusão num arquivo ou ainda para decisão sobre duplicidade. Face a isto, estas rotinas agora incluem uma função que apresenta o



cálculo de valores de escore mínimos e máximos para o conjunto de opções selecionado (Figura 5).



Figura 5

Cálculo de escores segundo parâmetros informados.

### DUPPLICIDADE

O relacionamento interno de bases constitui um caso especial da técnica de relacionamento, que tem como objetivo a identificação de registros duplicados internamente em uma base de dados. Para que este objetivo seja alcançado, cria-se um arquivo (denominado **mestre**) que contém apenas ocorrências únicas de registros identificados por um conjunto de campos-chave, extraído de um ou mais arquivos (chamados de **movimento**). Essa rotina foi, então, implementada apresentando como características: (1) relacionamento feito internamente à base (sem necessidade de duplicação da base); (2) padronização, blocagem e pareamento feitos em um único processo; (3) criação de campo-chave que é atualizado no arquivo reduzido (sem duplicidades) e no arquivo de entrada.

A Figura 6 apresenta o campo-chave (denominado nesse exemplo de “Único”) atualizado no arquivo de entrada (nesse exemplo, SINANEX.DBF – note-se que os nomes mostrados na figura são todos fictícios: a base utilizada na geração das telas mostradas neste artigo não contém informação de pacientes reais).

A tela 'Visualiza Arquivos' exibe o conteúdo do arquivo 'C:\Programas\ReCLink\Dados\SINANEX.DBF'. A tabela resultante possui as seguintes colunas e dados:

IDANEXO	ARCEITE	OVERSET	NOME	IDENF	UNICO
			IVO	200608141	20060814143004414
			LUIZ	200608141	20060814143004424
			VALDIR	200608141	20060814143004434
			ANA MORE	200608141	20060814143004444
			JOAO	200608141	20060814143004454
			GUERTE	200608141	20060814143004464
			ANA LUCIA	200608141	20060814143004474
			CARLOS	200608141	20060814143004484
			CARLOS	200608141	20060814143004494
			GERALDO	200608141	20060814143004504
			CARLOS	200608141	20060814143004514
			CARLOS	200608141	20060814143004524

Na barra de status inferior, é indicado 'Total de registros: 420' e há um botão 'Retorna'.

Figura 6

Campo-chave criado pela rotina de duplicidades.

#### CONSIDERAÇÕES FINAIS

O projeto de pesquisa que deu origem ao artigo foi aprovado pelo comitê de ética em pesquisa do IMS/UERJ (número: 17/2005).

A nova versão estará disponível em breve para *download* no *site* do RecLink (<http://paginas.terra.com.br/educacao/kencamargo/RecLinkII.html>), também gratuitamente para todos os interessados. Continua-se a contar com a colaboração destes para o contínuo aprimoramento do programa.

\* Agradecimentos: Os autores agradecem ao Programa Nacional de DST/AIDS e ao CNPq (projeto: 471562/2004-1) o financiamento para o desenvolvimento da nova versão, e a todos os usuários do programa, o contínuo estímulo.

#### REFERÊNCIAS BIBLIOGRÁFICAS

BORLAND INTERNATIONAL INC. *Borland C++ Developer Studio*. Scotts Valley, California, USA, 2006. Disponível em: <<http://www.borland.com>>. Acesso em: 01 jun. 2005.

CAMARGO JR., K. R.; COELI, C. M. Reclink: aplicativo para o relacionamento de banco de dados implementando o método probabilistic record linkage. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 16, n. 2, p. 439 - 447, 2000.

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. *Journal of the American Statistical Association*, v. 64, n. 328, p. 1183 - 1210, 1969.

## INSTRUÇÕES PARA OS COLABORADORES

Os *Cadernos Saúde Coletiva* publicam trabalhos inéditos considerados relevantes para a área de Saúde Coletiva.

SERÃO ACEITOS TRABALHOS PARA AS SEGUINTE SEÇÕES:

**Artigos** (resultantes de pesquisa de natureza empírica, experimental ou conceitual, ou ensaios teóricos e/ou de revisão bibliográfica crítica sobre um tema específico; máximo de 25 páginas); **Debate** (a partir de apresentações orais em eventos científicos, transcritos e sintetizados; máximo de 20 páginas); **Notas** (relatando resultados preliminares ou parciais de pesquisas em andamento; máximo de 5 páginas); **Opiniões** (opiniões sobre temas ligados à área da Saúde Coletiva, de responsabilidade dos autores, não necessariamente refletindo a opinião dos editores; máximo 5 páginas); **Cartas** (curtas, com críticas a artigos publicados em números anteriores; máximo de 2 páginas); **Resenhas** (resenhas críticas de livros ligados à Saúde Coletiva; máximo de 5 páginas); **Teses** (resumo de trabalho final de Mestrado, Doutorado ou Livre-Docência, defendidos nos últimos dois anos; com nome do orientador, instituição, ano de conclusão, palavras-chave, título em inglês, *abstract* e *key words*; máximo 2 páginas).

### APRESENTAÇÃO DOS MANUSCRITOS:

Serão aceitos trabalhos em português, espanhol, inglês ou francês. Os originais devem ser submetidos em três vias em papel, juntamente com o respectivo disquete (formato .doc ou .rtf), com as páginas numeradas. Em uma folha de rosto deve constar: Título em português e em inglês, nome(s) do(s) autor(es) e respectiva qualificação (vinculação institucional e título mais recente), endereço completo do primeiro autor (com CEP, telefone e *e-mail*) e data do encaminhamento. O artigo deve conter título do trabalho em português, título em inglês, resumo e *abstract*, com palavras-chave e *key words*. As informações constantes na folha de rosto não devem aparecer no artigo. Sugere-se que o artigo seja dividido em sub-itens. Os artigos serão submetidos a no mínimo dois pareceristas, membros do Conselho Científico dos Cadernos ou eventualmente *ad hoc*. O Conselho Editorial dos *Cadernos Saúde Coletiva* enviará carta resposta informando da aceitação ou não do trabalho.

A aprovação dos textos implica a cessão imediata e sem ônus dos direitos autorais de publicação nesta revista, a qual terá exclusividade de publicá-los em primeira mão. O autor continuará a deter os direitos autorais para publicações posteriores.

Caso a pesquisa que der origem ao artigo encaminhado aos Cadernos tenha sido realizada em seres humanos, será exigido que esta tenha obtido parecer favorável de um Comitê de Ética em Pesquisa em Seres Humanos, devendo o artigo conter a referência a este consentimento, estando citado qual CEP o concedeu, e cabendo a responsabilidade pela veracidade desta informação exclusivamente ao autor do artigo.

- **Formatação:** Os trabalhos devem estar formatados em folha A4, espaço duplo, fonte Arial 12, com margens: esq. 3,0 cm, dir. 2,0 cm, sup. e inf. 2,5 cm. Apenas a primeira página interna deverá conter o título do trabalho; o título deve vir em negrito e os subtítulos em versaleta (PEQUENAS CAPITAIS) e numerados; palavras estrangeiras e o que se quiser destacar devem vir em itálico; notas explicativas, caso existam, deverão vir no pé de página; as citações literais com menos de 3 linhas deverão vir entre aspas dentro do corpo do texto; as citações literais mais longas deverão vir em outro parágrafo, com recuo de margem de 3 cm à esquerda e espaço simples. Todas as citações deverão vir seguidas das respectivas referências.

- Ilustrações: o número de quadros e/ou figuras (gráficos, mapas etc.) deverá ser mínimo (máximo de 5 por artigo, salvo exceções, que deverão ser justificadas por escrito em anexo à folha de rosto). As figuras poderão ser apresentadas em nanquim ou produzidas em impressão de alta qualidade, e devem ser enviadas em folhas separadas e em formato .tif. As legendas deverão vir em separado, obedecendo à numeração das ilustrações. Os gráficos devem ser acompanhados dos parâmetros quantitativos utilizados em sua elaboração, na forma de tabela. As equações deverão vir centralizadas e numeradas sequencialmente, com os números entre parênteses, alinhados à direita.

- Resumo: todos os artigos submetidos em português ou espanhol deverão ter resumo na língua principal (Resumo ou *Resumen*, de 100 a 200 palavras) e sua tradução em inglês (*Abstract*); os artigos em francês deverão ter resumo na língua principal (*Résumé*) e em português e inglês. Deverão também trazer um mínimo de 3 e um máximo de 5 palavras-chave, traduzidas em cada língua (*key words*, *palabras clave*, *mots clés*), dando-se preferência aos Descritores para as Ciências da Saúde, DeCS (a serem obtidos na página <http://decs.bvs.br/>).

- Referências: deverão seguir a Norma NBR 6023 AGO 2000 da ABNT. No corpo do texto, citar apenas o sobrenome do autor e o ano de publicação, seguido da página no caso de citações (Sobrenome, ano: página). No caso de mais de dois autores, somente o sobrenome do primeiro deverá aparecer, seguido da expressão latina '*et al.*'. Todas as referências citadas no texto deverão constar nas REFERÊNCIAS BIBLIOGRÁFICAS ao final do artigo, em ordem alfabética, alinhadas somente à esquerda, pulando-se uma linha de uma referência para outra, constando-se o nome de todos os autores. No caso de mais de uma obra do mesmo autor, este deve ser substituído nas referências seguintes à primeira por um traço e ponto. Não devem ser abreviados títulos de periódicos, livros, locais, editoras e instituições.

Seguem exemplos de, respectivamente, artigo de revista científica impressa e veiculado via internet, livro, tese, capítulo de livro e trabalho publicado em anais de congresso (em casos omissos ou dúvidas, referir-se ao documento original da Norma adotada):

ESCOSTEGUY, C. C.; MEDRONHO, R. A.; PORTELA, M. C. Avaliação da letalidade hospitalar do infarto agudo de miocárdio do Estado do Rio de Janeiro através do uso do Sistema de Informações Hospitalares/SUS. *Cadernos Saúde Coletiva*. Rio de Janeiro, v.7, n.1, p. 39-59, jan./jul. 1999.

PINHEIRO, R.; TRAVASSOS, C. Estudo da desigualdade na utilização de serviços de saúde por idosos em três regiões da cidade do Rio de Janeiro. *Cadernos de Saúde Pública*. Rio de Janeiro, v.15, n.3, set. 1999. Disponível em: <<http://www.scielo.org/cgi-bin/wxis.exe/iah/>>. Acesso em: 2 jan. 2005.

ROSEN, G. *Uma história da Saúde Pública*. Rio de Janeiro: Abrasco. 1994. 400p.

TURA, L. F. R. *Os jovens e a prevenção da AIDS no Rio de Janeiro*. 1997. 183p. Tese (Doutorado em Medicina) - Faculdade de Medicina. UFRJ, Rio de Janeiro.

BASTOS, F. I. P.; CASTIEL, L. D. Epidemiologia e saúde mental no campo científico contemporâneo: labirintos que se entrecruzam? In: AMARANTE, P. (Org.) *Psiquiatria social e reforma psiquiátrica*. Rio de Janeiro: Fiocruz, 1994. p. 97-112.

GARRAFA, V.; OSELKA, G.; DINIZ, D. Saúde Pública, bioética e equidade. In: CONGRESSO BRASILEIRO DE SAÚDE COLETIVA, 5., 1997, Águas de Lindóia. *Anais*. Rio de Janeiro: Abrasco, 1997. p. 59-67.

# cadernos Saúde Coletiva

NESC • UFRJ



## ASSINATURA

Anual individual: R\$ 40,00

Anual institucional: R\$ 90,00

Número avulso: R\$ 20,00

Preencha este formulário de forma legível e envie-o anexando cheque nominal ou comprovante de depósito em favor da Fundação Universitária José Bonifácio (Banco do Brasil, ag.0287-9, c/c 7333-4, especificando na guia de depósito: depósito identificado - Apostila 9201-0), no valor correspondente, para a SECRETARIA DOS CADERNOS SAÚDE COLETIVA.

Este formulário também pode ser preenchido via internet na página: <http://www.nesc.ufrj.br>, ou enviado por e-mail para [cadernos@nesc.ufrj.br](mailto:cadernos@nesc.ufrj.br)

☐ Individual    ☐ Institucional    ☐ Número avulso: v.\_\_\_\_, n.\_\_\_\_

Nome: \_\_\_\_\_

Instituição: \_\_\_\_\_

Profissão: \_\_\_\_\_

Endereço: \_\_\_\_\_

Bairro: \_\_\_\_\_ CEP: \_\_\_\_\_

Cidade: \_\_\_\_\_ UF: \_\_\_\_\_

Tel.: (0xx\_\_ ) \_\_\_\_\_ Fax: \_\_\_\_\_

E-mail: \_\_\_\_\_

Local e data: \_\_\_\_\_

Assinatura: \_\_\_\_\_

