

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.

REVISÃO SISTEMÁTICA SOBRE ENCADEAMENTO OU *LINKAGE* DE BASES DE DADOS SECUNDÁRIOS PARA USO EM PESQUISA EM SAÚDE NO BRASIL

Systematic review on usage of large databases data linkage in research in public health in Brazil

João Paulo Lyra da Silva¹, Cláudia Travassos²,
Miguel Murat de Vasconcellos³, Linair Maria Campos⁴

RESUMO

Objetivo: Traçar um panorama da utilização da técnica de encadeamento de arquivos na área de Saúde Pública no Brasil de modo a gerar subsídios para seu aprimoramento e melhor emprego. **Método:** revisão sistemática de publicações de estudos sobre o Brasil na área de Saúde Pública que empregaram essa técnica. Incluídos estudos com Sistemas de Informação em Saúde do Sistema Único de Saúde (SIS-SUS) que objetivaram agregar, à informação disponível em um banco de dados sobre um indivíduo, outros dados sobre este mesmo indivíduo. Excluídos estudos em que o SIS-SUS foi encadeado somente a bases com agregação de dados de âmbito geográfico e de avaliação da qualidade dos dados a partir de fontes de dados dependentes. A unidade de análise foi o estudo. Neles foram analisados o desenho, os objetivos, as bases de dados encadeadas e em relação ao encadeamento os programas de computação utilizados, questões metodológicas relevantes, informações sobre sua validade e a qualidade dos dados nos bancos nele utilizados. **Resultados:** Identificadas 71 publicações referentes a 40 estudos. 28 estudos classificados como epidemiológicos. Predominaram o desenho de coorte, o uso de arquivos do SIM, do SINASC, algoritmos determinísticos e o tema mortalidade infantil, especialmente a neonatal. Nove estudos preocuparam-se com o planejamento e gestão dos sistemas e serviços de saúde. Seis objetivaram o desenvolvimento da técnica de encadeamento de arquivos e 6 analisar a qualidade dos dados das bases encadeadas. Trinta e um estudos informaram sobre a acurácia dos dados, 24 sobre a precisão, 34 sobre a completeza, 33 sobre a confiabilidade, e 31 sobre a atualidade dos dados nos bancos encadeados. Só foi possível obter informação sobre o percentual de pares verdadeiros não encontrados em 23 estudos e sobre revisão manual após o encadeamento automático em 28. **Discussão:** O crescimento do uso da técnica coincide com a implantação do SINASC e com a intensificação do uso das bases de dados do SIH-SUS na saúde coletiva. Mais temas poderiam ter sido abordados empregando-a para gerar, a baixo custo, conhecimento importante para o aprimoramento da Saúde. O *data warehouse* é ferramenta de grande valia para uma política abrangente de educação, comunicação e informação em Saúde.

PALAVRAS-CHAVE

Saúde pública, encadeamento de arquivos, uso de bases de dados secundários em pesquisa, informação em saúde

¹ Doutorando em Saúde Pública. ENSP/Fiocruz - e-mail: jplyra@uol.com.br

² PhD. Pesquisadora do DIS/CICT/Fiocruz

³ Doutor em Engenharia Biomédica. Pesquisador do DAPS/ENSP/Fiocruz

⁴ Mestre em Saúde Coletiva. NCE/UFRJ.

ABSTRACT

Objective: to delineate an overview of data linkage usage in Public Health in Brazil in order to generate subsidies for its improvement and better employment. **Method:** systematic revision of studies on Brazil in the area of Public Health in which this technique was used. Studies based on Health Information Systems of the Brazilian Unified Health System (SIS-SUS) were selected if the data linkage process aimed to add data on an individual to the information already available on such individual in a data base. Studies in which the SIS-SUS was linked only with bases aggregating data of geographic scope and studies on the evaluation of data quality based on dependent data were excluded. The unit of analysis was the study. Each study was analyzed for its design, its objectives, the linked databases and, as regards the linkage process, the computer programs used, relevant methodological issues and information on the reliability and the quality of the data in the banks used. **Results:** 71 publications referring to 40 studies were identified. 28 studies were classified as epidemiologic. Cohort design, the use of infant mortality systems files and of live birth systems files, deterministic algorithms and the issue of infant mortality, especially neonatal mortality, were all found to predominate. Nine studies addressed health systems and services planning and management. Six aimed at the development of the data linkage technique and 6 aimed at analyzing the quality of the data of the linked bases. On 31 studies the accuracy, on 24 the precision, on 34 the completeness, on 33 the trustworthiness, and on 31 the currentness of data on the linked files was informed. Information on the percentage of true pairs not matched could only be obtained in 23 studies and on manual revision after automatic data linkage in only 28 studies. **Discussion:** The escalation of the use of the technique coincides with the implementation of the live birth system and the intensification in the use of databases of the hospital admissions system in collective health. Many other issues could have been approached through its use in order to generate, at low cost, important knowledge for the improvement of Health. Data warehouse is an invaluable tool for a comprehensive policy of Health education, communication and information.

KEY WORDS

Public health, data linkge, large databases research, health information

INTRODUÇÃO

Nos últimos tempos, verifica-se uma acelerada expansão das tecnologias de informação. Associadas às transformações ocorridas na maquinaria, observam-se quatro etapas: o aparecimento e disseminação dos grandes computadores (*Mainframes*); os minicomputadores; a popularização dos computadores pessoais; e o crescimento da Internet. Atualmente, sistemas antigos de grande porte convivem com sistemas tecnologicamente atualizados e em constante desenvolvimento.

A migração de uma tecnologia em obsolescência para uma nova tecnologia em fase de adoção pode ser custosa e arriscada. A integração de sistemas antigos com novos sistemas pode atenuar os riscos de uma migração abrupta, ou mesmo permitir que sistemas antigos integrados, que bem atendam às necessidades de seus usuários, não precisem ser reimplementados com tecnologia atualizada.

A interoperabilidade, efetiva interconexão entre distintos sistemas de computador, bancos de dados ou redes para apoio ao intercâmbio de dados ou à computação distribuída, visa tal integração.

A interoperabilidade traduz-se na capacidade de diferentes máquinas, programas e linguagens trabalharem de forma compartilhada. Permite troca de dados e chamadas de rotinas de modo transparente para o usuário. No entanto, ainda que desejável, raramente uma interoperabilidade plena é atingida. Costuma, geralmente, ser dificultada por barreiras decorrentes das diferenças não só nas plataformas das máquinas, mas também dos programas de computação, incluindo sistemas operacionais, modelos de programação e de dados (Pacheco & Kern, 2001).

O setor Saúde realiza volumoso número de procedimentos, utiliza grande variedade de equipamentos diagnósticos e terapêuticos e emprega um enorme contingente de recursos humanos. Nesse ambiente de grandes magnitudes e em permanente evolução, foram desenvolvidos em diferentes momentos vários Sistemas de Informações em Saúde. Essa multiplicidade implica em fragmentação da informação. Para evitar duplicidade de dados, coleta de dados desnecessários e sobrecarga dos profissionais de saúde, com possível impacto negativo na credibilidade e confiabilidade dos Sistemas de Informação, deve-se, idealmente, buscar sua articulação e integração (Leão *et al.*, 2004). No Brasil, os registros epidemiológicos e administrativos do Sistema Único de Saúde – SUS – cobrem eventos distintos; por exemplo: nascidos vivos (SINASC); óbitos (SIM); agravos de notificação compulsória (SINAN); internações hospitalares (SIH-SUS) e procedimentos ambulatoriais de alto custo e alta complexidade pagos pelo SUS (APAC-SIA).

O encadeamento de arquivos (*linkage*) é a integração de informações de duas fontes de dados independentes. No encadeamento, os registros de cada uma das fontes, tidos como relacionados a um mesmo indivíduo, são pareados (*matched*) de forma a poderem ser tratados como um registro único para aquele indivíduo. O pareamento é empregado para se juntar registros correspondentes ao mesmo indivíduo de dois ou mais arquivos ou duplicatas de registros dentro de um mesmo arquivo. O encadeamento de arquivos por pareamento pode permitir uma interoperabilidade entre Sistemas de Informações, embora de modo limitado. Representa uma interoperabilidade que demanda um esforço mínimo de padronização.

Winkler (2001), do Bureau do Censo dos Estados Unidos da América – EUA, aponta existirem pelo menos quatro situações em que é indicada a utilização do encadeamento de arquivos. A saber: na estimativa de uma população pela técnica de captura-recaptura; na atualização e eliminação de duplicidades em um arquivo que contenha o registro chave de sujeitos de um inquérito; na agregação de dois

arquivos, sendo pelo menos um deles de cunho estatístico, visando análises do microdado; e na re-identificação, para garantir a confidencialidade de microdados para uso público. A re-identificação promove a proteção da privacidade do cidadão, por exemplo, no trânsito de arquivos entre diferentes agências do governo, impedindo sua identificação.

Na maior parte dos países do mundo, o órgão oficial de estatística responde pela produção de pesquisas econômicas, sociais e demográficas, além de coordenar a produção de informação realizada por outros órgãos. No Brasil, todos os órgãos e entidades da administração pública direta e indireta de todas as esferas de governo e entidades privadas que recebem subvenção ou auxílio para a produção de informações estatísticas formam o Sistema Estatístico Nacional (SEN). A função de coordenador do SEN é exercida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) (Silva, 2005).

O encadeamento de arquivos, segundo Fellegi (1997), tem sido de particular interesse nos países que, na tensão entre direitos individuais e as demandas do Estado, favorecem o equilíbrio nessas relações. Nestes casos, o encadeamento de arquivos acontece para realizar uma tarefa difícil na medida em que exista um consenso social que ela deva permanecer difícil em prol da privacidade das pessoas.

No setor saúde o encadeamento de arquivos pode ser usado para melhorar a qualidade dos dados; para permitir a reutilização de bases de dados secundários para finalidades diferentes daquelas para as quais o banco foi criado; e para diminuir os custos na aquisição de dados para pesquisa. Fair (1997) lista alguns exemplos do emprego do encadeamento de arquivos em estudos na Saúde Pública: no acompanhamento temporal de eventos vitais – nascimento e óbitos ou de eventos mórbidos específicos; na construção, manutenção e uso de registros administrativos de saúde; na recuperação de dados sobre a história clínica de pacientes; em estudos ecológicos; e no exame de fatores que influenciam o uso e o processo, os custos dos cuidados de saúde.

Existem vários algoritmos para se encadear arquivos de dados. Em conjuntos de dados que têm um identificador único das pessoas, ou uma chave, o problema de encadeamento torna-se simples: tudo o que se exige é uma operação em programas de análise de dados. A associação de arquivos com chave comum é usualmente suportada de forma direta por sistemas gerenciadores de bancos de dados (SGBD). Porém, na maioria dos casos, não há identificador único e algoritmos de encadeamento mais sofisticados precisam ser aplicados. Esses algoritmos podem ser separados em dois grupos não excludentes entre si: algoritmos determinísticos ou baseados em regras, em que conjuntos de regras freqüentemente muito complexas são usados para classificar pares de registros como vinculados, isto é, relativos à mesma pessoa ou entidade, ou como não vinculados; e algoritmos

probabilísticos, em que modelos estatísticos são usados para classificar pares de registros. Algoritmos probabilísticos podem, por sua vez, ser subdivididos naqueles baseados na teoria probabilística clássica de encadeamento de arquivos, como a desenvolvida por Fellegi e Sunter (1969), e em algoritmos mais recentes que usam entropia máxima, redes bayesianas, agrupamentos (*clustering*) e outras técnicas de aprendizagem de máquina (Christen *et al.*, 2004; Coutinho, 2004; Gu *et al.*, 2003, Winkler, 2003).

Este artigo objetiva traçar um panorama da utilização do encadeamento de arquivos na área de Saúde Pública no Brasil, quais funções tem desempenhado e identificar as estratégias de encadeamento mais usuais empregadas. Busca-se gerar subsídios para aprimoramento do emprego do encadeamento na Saúde Pública.

MÉTODO

Realizou-se uma revisão sistemática da literatura em Saúde Pública sobre estudos que empregaram a técnica de encadeamento de arquivos. Nessa revisão, selecionaram-se apenas estudos em que foram encadeados arquivos de pelo menos um dos Sistemas de Informação em Saúde do Sistema Único de Saúde (SIS-SUS).

As ferramentas de busca foram o MEDLINE e o LILACS, por intermédio da pesquisa avançada simultânea das bases de dados da Biblioteca Virtual em Saúde – BVS – da BIREME, e os bancos de teses e dissertações da Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior – CAPES e da BVS – Saúde Pública. O período focado neste estudo foi até 2005.

Na BVS foram pesquisados os sintagmas: relação, inter-relacionamento, compatibilização, encadeamento, vinculação, concatenamento, comparação, cruzamento – e seus correlatos (por exemplo: substituiu-se relação por relacionar, relacionado, relacionados, relacionada, relacionadas, relacionamento) – associados aos termos sistema, dados e arquivos. Além dessas expressões, utilizaram-se os termos captura-recaptura e captura e recaptura. Os sintagmas e termos foram buscados com o uso do recurso do motor de busca que realiza a pesquisa nos idiomas português, inglês, espanhol, alemão, francês e italiano. A pesquisa ateuve-se a estudos sobre o Brasil. A partir dos artigos, dissertações e teses selecionados foram procuradas outras publicações pela produção de seus autores, através dos índices de autores do LILACS, do MEDLINE; no sistema de currículos da plataforma Lattes; e nas referências bibliográficas indicadas.

Foram incluídos apenas estudos em que o encadeamento objetivou agregar, à informação disponível em um banco de dados sobre um indivíduo, outros dados sobre esse mesmo indivíduo, ou para estudos específicos ou para estimativa de uma população por captura-recaptura. Foram excluídos estudos ecológicos ou espaciais em que o SIS-SUS foi encadeado somente a bases com agregação de

dados de âmbito geográfico e estudos voltados para avaliação da qualidade dos dados a partir de fontes de dados dependentes, como, por exemplo, os do SIH-SUS a partir de dados do prontuário.

Publicações referentes a um mesmo estudo foram agrupadas de forma que a unidade de análise foi o estudo. Em cada um deles foram descritos o desenho do estudo, seus objetivos, as bases de dados encadeadas, os programas de computação utilizados no encadeamento e questões metodológicas relevantes relativas ao encadeamento. Analisaram-se também as informações fornecidas pelos autores sobre a validade do encadeamento e a qualidade dos dados nos bancos utilizados.

A partir de seus objetivos os estudos foram classificados como: relacionados ao desenvolvimento da técnica de encadeamento de arquivos; voltados para análise da qualidade dos dados das bases encadeadas; relacionados à área de planejamento, gestão e de avaliação de sistemas e serviços de saúde; e epidemiológicos. Os relacionados ao desenvolvimento da técnica de encadeamento de arquivos referem-se aos que objetivaram obter um produto ou processo que contribua para o aperfeiçoamento dessa técnica. Os voltados para análise da qualidade dos dados das bases encadeadas objetivaram analisar se os dados das bases encadeadas estão de acordo com as necessidades do usuário para a produção de informações confiáveis, de adequados valor e utilidade. Os relacionados à área de planejamento, gestão e de avaliação de sistemas e serviços de saúde objetivaram analisar normas, processos e condutas no âmbito dos sistemas e serviços de saúde. Os estudos epidemiológicos objetivaram analisar a distribuição quantitativa dos fenômenos de saúde/doença em populações humanas e seus fatores condicionantes, ou a avaliação da eficácia de intervenções realizadas no âmbito da saúde pública. O mesmo estudo pôde ser enquadrado em mais de uma categoria.

Quanto à validade do encadeamento foram consideradas as informações do percentual de perdas, que corresponde ao percentual de pares verdadeiros não encontrados, e as causas mais freqüentes dessas perdas; e as informações da realização ou não de revisão manual depois do encadeamento automático, com algoritmos probabilísticos ou determinísticos. Em relação à qualidade dos dados nos bancos utilizados no encadeamento, buscaram-se nos estudos informações sobre acurácia; precisão; completeza; confiabilidade; e atualidade. A acurácia – cuidado, exatidão, grau de conformidade com o padrão estabelecido – diz respeito ao grau de correção sintática do dado e também à sua semântica. A precisão diz respeito ao grau que o valor do dado corresponde ao valor real, dado que a diferença ocorre devido a aproximações propositais e não a erros, como no caso da falta de acurácia. A completeza diz respeito ao grau de existência da informação no banco de dados e à forma de tratamento de valores nulos. A confiabilidade, neste caso, expressa o grau de confiança que usuários têm nos dados e é um indicador

subjetivo. Atualidade diz respeito à época em que a informação foi disponibilizada. Deve-se lembrar que o percentual de perdas depende diretamente da qualidade do dado nos arquivos utilizados.

RESULTADOS

Foram identificadas setenta e uma publicações referentes a quarenta estudos. Vinte dissertações de mestrado e doze teses de doutorado produziram vinte e oito artigos científicos que, junto com mais nove artigos, foram publicados em treze periódicos. As duas outras foram publicadas como comunicados em congressos (Tabela 1). As revistas que mais os publicaram, dez cada uma, foram os *Cadernos de Saúde Pública* e a *Revista de Saúde Pública*, seguidas pela *Revista Brasileira de Epidemiologia*, com quatro artigos. A *Revista Brasileira de Saúde Materno Infantil*, o *Brazilian Journal of Infectious Diseases* e a *Revista Brasileira de Ginecologia e Obstetria* publicaram dois artigos e o *Informe Epidemiológico do SUS*, a *Epidemiologia e Serviços de Saúde*, o *Journal of Biosocial Science*, o *Jornal de Pediatria*, o *Jornal Brasileiro de Pneumologia*, o *Cadernos Saúde Coletiva* e a *Revista Latino-Americana de Enfermagem*, um cada. Os autores, oitenta e nove profissionais, eram vinculados a dezessete instituições de ensino e pesquisa, sendo três estrangeiras, e a quatro serviços de saúde: um hospital de ensino, duas secretarias municipais de saúde e uma secretaria estadual de saúde. A origem institucional dos autores concentrou-se em instituições de ensino e pesquisa localizados no Rio de Janeiro e em São Paulo.

Quanto aos objetivos, seis estudos foram classificados como de desenvolvimento da técnica de encadeamento de arquivos, sendo duas análises do uso do encadeamento de arquivos do SINASC com os do SIM (Almeida & Mello Jorge, 1996; Fernandes, 1997); uma análise exploratória sobre as potencialidades dos bancos do movimento do registro civil para o estudo da mortalidade infantil por raça/cor em uma coorte de nascidos vivos no estado de São Paulo (Cunha, 2001); dois voltados para desenvolver algoritmos para encadeamento de arquivos do SIH-SUS (Portela *et al.*, 1997) e da APAC-SIA (Gomes Júnior & Almeida, 2004); e uma análise da factibilidade da implantação de um sistema de vigilância do *diabetes mellitus* na população idosa (Coeli, 1998). A partir de necessidades surgidas no decorrer do estudo de Coeli (1998), desenvolveu-se um sistema informatizado, o RecLink® (Camargo Jr. & Coeli, 2000), e publicou-se um procedimento para utilizá-lo, testada a eficiência de diferentes esquemas de blocagem (Coeli & Camargo Jr., 2002). O estudo de Machado (2002) produziu uma visão panorâmica, com base em revisão bibliográfica, dos procedimentos de encadeamento de arquivos, com foco na mortalidade infantil (Machado, 2004); e uma rotina de padronização na utilização do RecLink® (Machado & Hill, 2004).

Tabela 1
Publicações encontradas e características dos estudos.

Estudo	Publicações	Tipo	SIS	Softwares	Descrição / Desenho do estudo
1. ALMEIDA	ALMEIDA, 1994 ALMEIDA & MELLO JORGE, 1996 ALMEIDA & MELLO JORGE, 1998	Desenvolvimento da técnica e epidemiológico	SIM/SINASC		Coorte de nascidos vivos no Município de Santo André, SP, no 1º semestre de 1992 e os óbitos neonatais entre 1/1/92 a 27/7/92, dela derivados.
2. FERNANDES	FERNANDES, 1997	Desenvolvimento da técnica	SIM/SINASC	Manual	Experiência metodológica de concatenamento de informações sobre óbitos e nascimentos do Distrito Federal - 1989-1991.
3. NORONHA	NORONHA <i>et al.</i> , 1997	Análise da qualidade dos dados das bases encadeadas	SIM/SINASC	Dbase 5.0 Eplinfo 6.04	Análises de concordância de dados em relação a DN como informação original - exceto a informação sobre o sexo da criança a óbitos neonatais do SIM no Rio de Janeiro, RJ, ocorridos no segundo semestre de 1996 com nascimento de 1 de julho a 31 de dezembro do mesmo ano.
4. ACCIOLY	ACCIOLY, 1997	Epidemiológico	SIM/SINASC	Eplinfo 6.0 e SAS	Coorte de nascidos vivos e óbitos neonatais em Belo Horizonte 1993.
5. GOMES	GOMES, 1993 GOMES & SANTO, 1997	Epidemiológico	SIM/SINASC	Manual	Estudo transversal descritivo da mortalidade de crianças menores de um ano, Presidente Prudente, SP, de 1990 a 1992.
6. BOHLAND	BOHLAND, 1996 BOHLAND & MELLO JORGE, 1999	Epidemiológico	SIM/SINASC	Manual	Estudo transversal descritivo da mortalidade de menores de cinco anos Itararé, SP, 1993.
7. FERNANDES & ALMEIDA	ALMEIDA <i>et al.</i> , 2002 FERNANDES, 2002	Epidemiológico	SIM/SINASC	Manual	Coorte de recém-nascidos primeiro semestre de 1995 e óbitos neonatais em São Paulo capital.
8. CARVALHO	CARVALHO, 2003	Epidemiológico	SIM/SINASC	SPSS e Eplinfo 6.04	Coorte de nascidos vivos no Recife, 1999.
9. CUNHA	CUNHA, 2004 CUNHA, 2001	Desenvolvimento da técnica	SIM/SINASC	Access com comandos SQL	Coorte de nascidos vivos no Estado de São Paulo em 1997 e 1998.
10. MORAIS NETO 1	MORAIS NETO, 1996 MORAIS NETO & BARROS, 2000	Epidemiológico	SIM/SINASC	FoxPro, Eplinfo 6.03, SAS (catmod)	Coorte de nascidos vivos 1992, Goiânia, GO.

11. MORAIS NETO 2	MORAIS NETO, 2002	Epidemiológico	SIM/SINASC /i.g	FoxPro, EpiInfo 6.04	Coorte de nascidos vivos 1992 a 1996, Goiânia, GO.
12. GIGLIO	GIGLIO, 2003 GIGLIO <i>et al.</i> , 2005b GIGLIO <i>et al.</i> , 2005a	Epidemiológico	SIM/SINASC	FoxPro 5.0 EpiInfo 6.04 SPSS 11.5	Coorte de nascidos vivos e óbitos neonatais em Goiânia 2000.
13. LI	LI, 2000 LI <i>et al.</i> , 2002	Planejamento e gestão e epidemiológico	SIM/SINASC /i.g	EpiInfo 6.04	Vigilância à saúde e gestão da informação nos SILOS: uma proposta a partir do SINASC no Município de São José.
14. MACHADO	MACHADO, 2002 MACHADO & HILL, 2005 MACHADO, 2004 MACHADO & HILL, 2003 MACHADO & HILL, 2004	Epidemiológico e desenvolvimento da técnica	SIM/SINASC	ReLink	Coorte de nascidos vivos, 1998, São Paulo, SP.
15. MARTINS	MARTINS, 2002 MARTINS & VALASQUES, 2004	Epidemiológico	SIM/SINASC	EpiInfo 6.0	Coorte de nascidos vivos de 1997 a 1999, Montes Claros, MG.
16. ORTIZ FLORES	ORTIZ FLORES, 1999	Epidemiológico	SIM/SINASC	Dbase	Coorte de nascidos vivos, de jan a mar de 1993, no Estado de São Paulo.
17. PETERS	PETERS, 2001	Epidemiológico	SIM/SINASC	EpiInfo 6.04	Coorte de nascidos vivos de gestação única em Joinville, 1995.
18. PINHEIRO	PINHEIRO, 2003	Análise da qualidade dos dados das bases encadeadas e planejamento e gestão	SIM/SINASC		Coorte de nascidos vivos, Ilhéus, 2001.
19. ROSSO	ROSSO, 2003 WEIRICH <i>et al.</i> , 2005	Epidemiológico	SIM/SINASC	FoxPro 6.0 EpiInfo 6.04 SPSS 10.01	Recém-nascidos de zero a 28 dias de idade em UTIN, de nov 1999 até ou 2000 em Goiânia, GO.
20. HELENA	HELENA <i>et al.</i> , 2005	Epidemiológico	SIM/SINASC	Stata 8.0	
21. SARACENI 1	SARACENI, 2005 SARACENI & LEAL, 2003	Epidemiológico e planejamento e gestão	SIM/SINASC /SINAN	Excel, Dbase 4.0 ReLink, EpiInfo 6.04	

Estudo	Publicações	Tipo	SIS	Softwares	Descrição / Desenho do estudo
22. SARINHO	SARINHO, 1998 SARINHO <i>et al.</i> , 2001	Epidemiológico	SIM/SINASC	Epiinfo 6.0 SPSS 4.0	
23. SOARES	SOARES, 2003	Epidemiológico	SIM/SINASC	Epiinfo 6.04 Stata 7.0 Reclink	
24. CAVALCANTE	CAVALCANTE, 2003 CAVALCANTE <i>et al.</i> , 2004 CAVALCANTE <i>et al.</i> , 2005	Planejamento e gestão	SINASC/L/ SINAN	Manual Excel Epiinfo 6.04	
25. FERREIRA	FERREIRA, 1999 FERREIRA <i>et al.</i> , 2000 FERREIRA & PORTELA, 1999	Planejamento e gestão	SIH/SINAN	SAS	
26. SARACENI 2	SARACENI <i>et al.</i> , 2005	Epidemiológico	SIM/SINAN/ SICOM	Reclink, Epiinfo 6.04, SPSS9.0	
27. LEMOS	LEMOS, 1999 LEMOS & VALENTE, 2001	Epidemiológico e planejamento e gestão	SIM/SINAN	Dbase 4	
28. SELIG	SELIG <i>et al.</i> , 2004	Planejamento e gestão	SIM/SINAN	Reclink, Epiinfo 6.04	
29. BRUM	BRUM & KUPEK, 2005	Epidemiológico	SINAN/L/ USFM	Stata	

30. COELI 1	COELI, 1998 CAMARGO JR. & COELI, 2000 COELI & CAMARGO JR., 2002 COELI <i>et al.</i> , 2004	Planejamento e gestão e desenvolvimento da técnica	SIH/SIM/estat amb SUS	Reclink Reclink	1) Enc arquivos ambulatorial (363 registros), hospitalar (134 registros) e de mortalidade (291 registros), relativos aos registros dos pacientes com diagnóstico de diabetes = 740 registros. 2) Enc arquivo ambulatorial (363 registros), arquivo com os óbitos por diabetes em 1994 (190 registros), arquivo óbitos idosos residentes na AP 2.2 (ano 95 - 2.857 registros) e arquivo hospitalizações idosos residentes na cidade Rio de Janeiro (anos 94 e 95 - 79.039 registros) = 82.449 registros.
31. VEIRA	VEIRA, 2004	Epidemiológico	SIH/SIM	EpiInfo 6.04 Stata 5.0 Reclink	Sobrevida de idosos com 60 anos ou mais em 1998 após internação pelo SUS no Rio de Janeiro, RJ por Fratura Proximal de Fêmur.
32. MELLO	MELLO, 2003	Epidemiológico	SIH/SIM	Excel, Reclink	Coorte de casos de AVE do ano de 1998 no Est. RJ seguida por um período de um ano.
33. ANDRADE	ANDRADE, 1998 ANDRADE & MELLO-JORGE, 2000 ANDRADE & MELLO-JORGE, 2001a ANDRADE & MELLO-JORGE, 2001b	Epidemiológico e análise da qualidade dos dados das bases encadeadas	BOPW/ SIH/SIM/CAT	EpiInfo	Vítimas de acidentes de transporte terrestre entre 1 de janeiro a 30 de junho de 1996, em Londrina, PR.
34. MARTINS	MARTINS, 2004 MARTINS e ANDRADE 2005a MARTINS & ANDRADE, 2005b	Epidemiológico	Inquérito/SIM	Manual e EpiInfo 6.0	Estudo transversal menores de 15 anos de idade em Londrina, PR, vítimas de acidente ou violência atendidas em serviços de emergência/urgência ou internação em 2001 ou, ainda, que morreram sem assistência médica nesse ano, ou em prazo de até um ano após o acidente ou ato violento ocorrido em 2001.
35. PORTELA	PORTELA <i>et al.</i> , 1997	Desenvolvimento da técnica	SIH/SIH	SAS	AIHs 1 e 5 faturadas em 1995, no Rio de Janeiro, RJ, correspondeu a 10% do total de AIHs, re
36. GOMES JR.	GOMES JR. & ALMEIDA, 2004	Desenvolvimento da técnica	APAC/APAC	SAS 8.2	

Estudo	Publicações	Tipo	SIS	Softwares	Descrição / Desenho do estudo
37. PEPE	PEPE, 2002	Epidemiológico	CPIHP/SIH	SAS e RecLink	Internações a reinternações psiquiátricas realizadas no SUS, Est. RJ, entre ago. 1995 e dez. divididas em três subpopulações: de única internação, de reinternação e moradores do hospital.
38. COELI 2	COELI <i>et al.</i> , 2003	Epidemiológico e análise da qualidade dos dados das bases encadeadas	Inquérito de base domiciliar/SIH	RecLink e Stata 7.0	Entrevistas domiciliares em Duque de Caxias. Inform. de residentes sobre a ocorrência
39. SILVA	SILVA, 2003	Planejamento e gestão	CPIHP/SIH/SI-CAPS	Programa em Visual Basic com Soundex	
40. GOES	GOES, 2003	Planejamento e gestão e epidemiológico	SINAVSICOM	RecLink e Stata 7.0 e EpiInfo 6.04	

Seis estudos objetivaram analisar a qualidade dos dados das bases encadeadas. Noronha e colaboradores (Noronha *et al.*, 1997) fizeram uma comparação entre os sistemas de informações de mortalidade e de nascidos vivos para analisar o grau de concordância do preenchimento de dados comuns a eles e recuperar informações. Andrade (1998) estudou vítimas de acidentes de transporte terrestre ocorridos no primeiro semestre de 1996, em Londrina, Paraná. As fontes de dados foram os boletins de ocorrência policial, as fichas de atendimento em pronto-socorro, os laudos das Autorizações de Internação Hospitalar (AIHs), as Declarações de Óbito e outras fontes complementares, como as Comunicações de Acidentes de Trabalho (CATs), notícias publicadas na imprensa escrita e entrevistas com as vítimas ou familiares. Além de caracterizar as pessoas de acordo com o papel desempenhado no acidente e através de outras variáveis de importância epidemiológica, mediu a cobertura e fidedignidade das diversas fontes oficiais de dados. Coeli e colaboradores (Coeli *et al.*, 2003) analisaram vantagens e limitações potenciais do uso das bases de dados dos formulários da AIH e da metodologia do relacionamento probabilístico de registros, para a validação de relatos de utilização de serviços hospitalares em inquéritos domiciliares. Ferreira (1999) quantificou a subnotificação de casos de Aids, tendo como base as informações sobre as internações por procedimentos relacionados à Aids realizadas em unidades hospitalares vinculadas ao SUS e Lemos (1999) quantificou e caracterizou o sub-registro de casos de Aids que evoluíram para óbito. Pinheiro (2003) procurou medir a confiabilidade do SINASC e do SIM como fontes de dados para obtenção da probabilidade de morte neonatal.

Nove estudos preocuparam-se com o planejamento e gestão dos sistemas e serviços de saúde. Objetivaram propor uma melhor utilização da informação visando intensificar a capacidade dos sistemas locais de saúde para detectar as desigualdades geográficas na saúde (Gomes, 1994); estudar a factibilidade da implantação de um sistema de vigilância do diabetes *mellitus* na população idosa empregando fontes de dados sobre morbidade e mortalidade usualmente disponíveis em nosso meio com base no emprego da metodologia de captura-recaptura (Coeli, 1998; Coeli *et al.*, 2004); analisar, no âmbito de um hospital universitário, as vantagens do uso articulado do Sistema de Controle da Dispensação de Medicamentos do Município do Rio de Janeiro (SICOM) e do SINAN para a vigilância epidemiológica da Aids (Goes, 2003); estudar os óbitos por tuberculose para monitorar o programa de controle de tuberculose (Selig *et al.*, 2004); medir a efetividade de campanhas de eliminação da sífilis congênita (Saraceni, 2005); identificar gestantes infectadas pelo HIV e analisar as condutas profiláticas utilizadas (Cavalcante, 2003); avaliar a mortalidade hospitalar, custos no SUS e tempo de permanência em internações de longa permanência

(Portela *et al.*, 1997); avaliar o processo de desospitalização e conseqüente absorção da clientela hospitalar pelos Centros de Atenção Psicossocial (CAPS) (Silva, 2003); testar duas formas de identificação de um caso novo de câncer no Sistema de Autorização de Procedimentos de Alta Complexidade em Oncologia (APAC ONCO/SUS) (Gomes Jr. & Almeida, 2004).

Vinte e oito estudos foram classificados como estudos epidemiológicos. Destes, dezoito trataram de mortalidade infantil (Almeida, 1994; Bohland, 1996; Morais Neto, 1996; Accioly, 1997; Sarinho, 1998; Ortiz Flores, 1999; Li, 2000; Peters, 2001; Almeida *et al.*, 2002; Fernandes, 2002; Martins, 2002; Morais Neto, 2002; Carvalho, 2003; Giglio, 2003; Saraceni & Leal, 2003; Pinheiro, 2003; Rosso, 2003; Soares, 2003; Helena *et al.*, 2005). Doze destes dezoito enfocaram a mortalidade neonatal e seus fatores de risco (Almeida, 1994; Bohland, 1996; Morais Neto, 1996; Accioly, 1997; Sarinho, 1998; Ortiz Flores, 1999; Almeida *et al.*, 2002; Fernandes, 2002; Martins, 2002; Morais Neto, 2002; Giglio, 2003; Rosso, 2003; Helena *et al.*, 2005). Dois eram concernentes à mortalidade relacionada à AIDS (Lemos, 1999; Saraceni, 2005) e um ao uso combinado de arquivos – sobre dispensação de medicamentos antiretrovirais de alta potência e uso de serviços hospitalares – na vigilância epidemiológica da Aids (Goes, 2003). A prevalência da leptospirose (Brum & Kupek, 2005); a prevalência e o padrão de sobrevida no acidente vascular encefálico (Mello, 2003); a sobrevida dos pacientes idosos com fratura proximal de fêmur (Vieira, 2004); as características das vítimas, do atendimento, da lesão e do incidente em acidentes e violências em menores de 15 anos (Martins, 2004); a caracterização das vítimas de acidentes de transporte (Andrade, 1998); a descrição de variáveis demográficas, clínicas, hospitalares e socioeconômicas de pacientes psiquiátricos internados e reinternados e a análise de fatores associados a morar no hospital, à reinternação ou a ter tido uma única internação (Pepe, 2002); e a avaliação das potenciais vantagens e limitações do uso do SIH-SUS e da metodologia do relacionamento probabilístico de registros para a validação de relatos de utilização de serviços hospitalares em inquéritos domiciliares (Coeli *et al.*, 2003) foram os temas dos outros sete estudos.

Em relação aos desenhos dos estudos, vinte são estudos de coorte (Almeida, 1994; Morais Neto, 1996; Accioly, 1997; Andrade, 1998; Ortiz Flores, 1999; Cunha, 2001; Peters, 2001; Fernandes, 2002; Machado, 2002; Morais Neto, 2002; Martins, 2002; Pepe, 2002; Carvalho, 2003; Giglio, 2003; Mello, 2003; Silva, 2003; Rosso, 2003; Soares, 2003; Vieira, 2004; Helena *et al.*, 2005); um estudo de caso-controle aninhado a uma coorte (Almeida *et al.*, 2002; Fernandes, 2002); dois de caso-controle (Saraceni, 2002; Sarinho, 1998); e oito estudos seccionais (Gomes, 1994; Bohland, 1996; Ferreira, 1999; Lemos, 1999; Cavalcante, 2003; Goes, 2003; Martins, 2004; Selig *et al.*, 2004).

A metodologia de captura-recaptura foi utilizada em dois estudos: na estimativa da população de idosos diabéticos em área da cidade do Rio de Janeiro (Coeli, 1998) e no número de casos de leptospirose humana no distrito de saúde de Santa Maria, no Rio Grande do Sul (Brum & Kupek, 2005). Apenas um estudo objetivou utilizar o encadeamento de arquivos para armazenamento de dados em apoio à tomada de decisão (*data warehouse*) (Li, 2000).

Trinta estudos encadearam arquivos do SIM; vinte e três, arquivos do SINASC; nove, do SIH-SUS; e oito, do SINAN. Dois estudos utilizaram o SICOM e dois outros utilizaram o Censo da População de Internos em Hospitais Psiquiátricos do Rio de Janeiro (CPIHP).

Vinte e dois estudos encadearam arquivos do SIM aos do SINASC. Dois deles encadearam os do SIM e do SINASC a arquivos com informações geográficas e um encadeou os do SIM e do SINASC aos do SINAN. O SIM foi também encadeado ao SIH-SUS em quatro estudos. Em um deles os dois sistemas foram encadeados a estatísticas ambulatoriais de atendimento ao *diabetes mellitus* em unidades de saúde credenciadas pelo SUS e, em outro, a Boletins de Ocorrência da Polícia Militar e Comunicações de Acidente de Trabalho (CAT). Dois estudos encadearam o SIM ao SINAN e outros dois estudos encadearam o SIM a inquéritos. Além dos vinte e dois estudos que encadearam arquivos do SINASC aos do SIM, um estudo encadeou arquivo do SINASC aos do LACEN e do SINAN.

Como apontado anteriormente, quatro estudos encadearam o SIH-SUS ao SIM. Arquivos do SIH-SUS foram encadeados também a arquivos do CPIHP em dois estudos; aos de um inquérito populacional em um estudo; e aos do SINAN em outro. Um estudo encadeou arquivos do SIH-SUS a outros arquivos do SIH-SUS.

Arquivos do SINAN foram encadeados a arquivos do SIM em três estudos; a do SICOM em dois; a do LACEN em dois; a do SIH-SUS em um estudo.

Um algoritmo determinístico de encadeamento de arquivos foi empregado em trinta e dois estudos. Em todos os onze estudos nos quais um algoritmo probabilístico de encadeamento de arquivos foi utilizado, informou-se o uso do programa de computação RecLink® (Camargo Jr. & Coeli, 2000). Três estudos utilizaram algoritmos determinísticos e probabilístico.

No estudo de Almeida (1994), o algoritmo utilizado no encadeamento arquivos foi determinístico, primeiro usando as variáveis sexo e data de nascimento, sendo depois feita a busca manual pela variável nome da mãe. Nos estudos de Accioly (1997), Ortiz Flores (1999), Peters (2001), Fernandes (2002), Martins (2002), Moraes Neto (2002) e Pinheiro (2003) usou-se o mesmo procedimento acima. Em Fernandes (1997), o encadeamento foi feito manualmente baseado no nome da mãe. Cinco estudos indicaram que o encadeamento não foi automático, ou seja, foi manual (Tabela 1). Em Noronha (1997), fez-se encadeamento de arquivos com algoritmo

determinístico de modo automático, usando o nome da mãe, sem o recurso à codificação fonética. Nesse mesmo estudo, dez mães homônimas foram discriminadas pelo peso ao nascer e sexo do concepto. A codificação fonética, utilizada no RecLink® e no programa de computação empregado no estudo de Silva (2003), permite contornar as dificuldades impostas ao pareamento por possíveis diferenças de grafia de fonemas iguais ou semelhantes nos registros a serem pareados no encadeamento automático de arquivos. Em Cunha (2001), a coorte foi selecionada por encadeamento automático de arquivos com algoritmo determinístico sem o uso dos nomes das mães, mas com o uso da data do nascimento, sexo, local de moradia e idade da mãe. A autora desse estudo relata que registros com valores nulos foram excluídos e que, ao fim, foi feita a comparação de todas as outras variáveis disponíveis utilizando um algoritmo determinístico. O número da Declaração de Nascido Vivo (DN), presente no SIM e no SINASC foi utilizado por Moraes Neto (1996) e Carvalho (2003) em uma etapa inicial seguida do pareamento por outros campos como o nome da mãe e a data do nascimento em Carvalho (2003) e data de nascimento, sexo e bairro de residência e nome da mãe em Moraes Neto (1996).

Só foi possível obter informação sobre o percentual de perdas (pares verdadeiros não encontrados) de vinte e três estudos. Machado (2002) recomenda que pesquisas que utilizem encadeamento de arquivos automático usem estratégias de pareamentos subsequentes realizados a partir do primeiro encadeamento, seguidos de uma revisão manual para recuperar pares de forma eficaz. Vinte e oito estudos informaram ter realizado revisão manual após o encadeamento automático.

Quanto à aferição da qualidade dos dados nos bancos utilizados no encadeamento, trinta e um estudos informaram sobre a acurácia dos dados, vinte e quatro sobre a precisão, trinta e quatro sobre a completude, trinta e três sobre a confiabilidade, e trinta e um sobre a atualidade dos dados nos bancos encadeados.

Fontes secundárias não são projetadas especificamente para fins de pesquisa e seus dados não se caracterizam pela alta qualidade exigida para essa finalidade (Machado, 2002). Foram reportados como problemas para a qualidade dos dados: erros ortográficos e de digitação; dados importantes sem informação, como, por exemplo, data de nascimento e sexo (Portela *et al.*, 1997; Ferreira, 1999); incorreções, como, por exemplo, pacientes com o mesmo nome e mesmo número de prontuário com datas de nascimento muito díspares (Portela *et al.*, 1997); e validade do diagnóstico questionável (Coeli *et al.*, 2003). Nomes pouco comuns iguais e datas de nascimento diferentes podem ser identificados como par provável, e nomes e sobrenomes comuns iguais e datas de nascimento diferentes podem ser considerados homônimos (Ferreira, 1999).

No SINASC e no SIM, nas variáveis sobre condição social da mãe da criança, há uma tendência à sobre-estimação no caso dos nascimentos e à subestimação no caso dos óbitos (Fernandes, 1997). No estudo de acidentes e violências é necessário melhoria, tanto da cobertura dos registros policiais como da qualidade da informação médica no que se refere às circunstâncias do acidente (Andrade, 1998).

Dois estudos apontaram que o SINASC contribui decisivamente para corrigir informações contidas no SIM (Fernandes, 1997; Cunha, 2001). O uso do SINASC é mais eficiente do que a coleta de informações no cartório para a obtenção de dados incompletos no SIM (Fernandes, 1997).

DISCUSSÃO

Os quarenta estudos em que foram encadeados arquivos de pelo menos um dos SIS-SUS no período de onze anos demonstram uma importante e recente utilização da técnica. Esse crescimento coincide com a implantação do SINASC e com a intensificação do uso das bases de dados do SIH-SUS na saúde coletiva (Bittencourt *et al.*, 2006).

A grande maioria, vinte e oito estudos, foi classificada como epidemiológico. Predominaram o desenho de coorte, o uso de arquivos do SIM, do SINASC, algoritmos determinísticos e o tema mortalidade infantil, especialmente a neonatal. Além da mortalidade neonatal, foram abordados temas como doenças sexualmente transmissíveis, notadamente a Aids, doenças crônico-degenerativas, como o diabetes *mellitus*, o acidente vascular encefálico e a fratura proximal do fêmur, e a violência. A escolha dos temas reflete a preocupação com a qualificação da assistência e a emergência da violência como tema da Saúde Pública. Não obstante, muitos outros temas poderiam ser abordados com a utilização do encadeamento de arquivos de SIS-SUS o que possibilitaria maior crescimento de seu emprego e a geração, a baixo custo, de conhecimento importante para o aprimoramento da Saúde.

Somente seis estudos objetivaram analisar a qualidade dos dados das bases encadeadas. Mas, trinta e um estudos informaram sobre a acurácia dos dados, vinte e quatro sobre a precisão, trinta e quatro sobre a completeza, trinta e três sobre a confiabilidade, e trinta e um sobre a atualidade dos dados nos bancos encadeados. Só foi possível obter informação sobre o percentual de pares verdadeiros não encontrados em vinte e três estudos e sobre revisão manual após o encadeamento automático em vinte e oito. Embora a qualidade dos bancos interfira na validade do encadeamento, pode-se entender, por aí, ter havido maior preocupação com o primeiro aspecto.

Aspectos de qualidade de dados têm sido largamente estudados na área da Ciência da Computação (Pinho, 2001; Wand & Wang, 1996). Em relação à melhor forma de se associar indicadores de qualidade dos dados de um banco, encontram-se, na literatura, várias propostas. Podemos diferenciá-las pelo lugar onde o dado é avaliado e pelo nível de detalhamento da informação de qualidade. Como exemplos: quanto ao lugar – no arquivo original ou após um processo de limpeza; e quanto ao nível de detalhamento da informação de qualidade – se a qualidade refere-se a valores, instâncias de um dado específico, tal como uma data de nascimento, ou se ela se refere a elementos de dados, um registro do banco de dados, ou ainda, ao banco de dados como um todo (Amaral & Campos, 2004). Não está no escopo deste trabalho a discussão de tais técnicas, mas pode-se sugerir como boa prática, com base em Amaral (2003), que se considerem alguns critérios, maneiras de aferi-los, e que se procure fazê-lo associando-os ao nível mais detalhado ou mais crítico possível, quando forem aplicáveis, a uma instância de dados. Por exemplo, se uma pesquisa baseia-se fortemente nos dados de data de nascimento do paciente, então esse dado deve ter associado a si indicadores de qualidade. Não esquecendo que, por outro lado, há necessidade de indicadores que digam respeito ao banco de dados como um todo, como é o caso da abrangência dos dados pesquisados.

No encadeamento de arquivos com algoritmo probabilístico a confiabilidade reflete-se na probabilidade que os elementos confrontados concordem sobre um par ajustado verdadeiro e é análoga à sensibilidade. Deveria ser determinada por revisão manual inicial dos dados ou através de pesquisa prévia. O poder discriminatório reflete-se na probabilidade que os elementos confrontados concordem sobre um par não ajustado e é análogo à especificidade. A qualidade dos dados impacta muito a qualidade do encadeamento. Os algoritmos determinísticos representam estratégias mais fáceis e mais rápidas, mas podem resultar em vieses significativos em função de não pareamentos ou perdas por pares verdadeiros não encontrados. Muitos pares verdadeiros não encontrados podem resultar em maior viés que poucos pareamentos errôneos (Bigelow *et al.*, 1999; Mason, 2006).

Apenas nove estudos objetivaram o planejamento e gestão dos sistemas e serviços de saúde e um único estudo utilizou o encadeamento de arquivos para armazenamento de dados em apoio à tomada de decisão (*data warehouse*). O *data warehouse*, com a reidentificação do usuário para preservar sua privacidade, seria de grande utilidade nos Centros de Documentação, Informação, Comunicação e Educação em Saúde voltados para a democratização das informações, cuja criação foi proposta na 10ª Conferência Nacional de Saúde. Entre as funções desses centros estariam as de subsidiar os processos de formulação de políticas e o controle e a fiscalização das políticas e ações governamentais por parte da população (Pitta & Magajewski, 2000).

A produção, organização e divulgação de dados de interesse para a saúde, integrado com ações e programas de educação e comunicação em saúde, de forma a subsidiar não somente ações técnicas, administrativas e gerenciais, mas também ações de capacitação profissional, educação para a cidadania, pesquisa e acompanhamento das ações e serviços de saúde, planejamento participativo, bem como programas interinstitucionais é um horizonte desejável (Pitta & Magajewski, 2000). A integração dos SIS-SUS representa um passo a mais em sua direção.

REFERÊNCIAS BIBLIOGRÁFICAS

ACCIOLY, M. C. *Determinantes da mortalidade neonatal em Belo Horizonte 1993: subsídios para o planejamento da atenção materno-infantil*. 1997. Dissertação (Mestrado em Medicina) - Faculdade de Medicina. UFMG, Belo Horizonte.

ALMEIDA, M. F.; MELLO JORGE, M. H. P.; O uso da técnica de “Linkage” de sistemas de informação em estudos de coorte sobre mortalidade neonatal. *Revista de Saúde Pública*. São Paulo, v. 30, n. 2, p.141 - 147, 1996.

_____. Pequenos para idade gestacional: fator de risco para mortalidade neonatal. *Revista de Saúde Pública*. São Paulo, v. 32, n. 3, p. 217 - 224, 1998.

ALMEIDA, M. F.; NOVAES, M. D. H.; ALENCAR, G. P.; RODRIGUES, L. C. Mortalidade neonatal no Município de São Paulo: influência do peso ao nascer e de fatores sócio-demográficos e assistenciais. *Revista Brasileira de Epidemiologia*. v. 5, n. 1, p. 93 - 107, 2002.

ALMEIDA, M. F. *Mortalidade neonatal em Santo André, SP, Brasil*. 1994. 140p. Tese (Doutorado em Saúde Pública) - Faculdade de Saúde Pública. USP, São Paulo.

AMARAL, G. C. M.; CAMPOS, M. L. M. AQUAWARE: A Data Quality Support Environment for Data Warehousing. Simpósio Brasileiro de Bancos de Dados. Distrito Federal. *Anais/Proceedings*. UnB 2004, p. 121 - 133. Disponível em: <<http://www.lbd dcc.ufmg.br/sbbd2004/paper40.pdf>>. Acesso em: 14 jun. 2006.

AMARAL, G. C. M. *AQUAWARE: um ambiente de suporte à qualidade de dados em Data Warehouse*. 2003. 157p. Dissertação (Mestrado em Ciência da Computação) - Núcleo de Computação Eletrônica do Instituto de Matemática. UERJ, Rio de Janeiro.

ANDRADE, S. M.; MELLO JORGE, M. H. P. Características das vítimas por acidentes de transporte terrestre em município da região sul do Brasil. *Revista de Saúde Pública*. São Paulo, v. 34, n. 2, p. 149 - 156, 2000.

_____. Acidentes de transporte terrestre em município da Região Sul do Brasil. *Revista de Saúde Pública*. São Paulo, v. 35, n. 3, p. 318 - 320, 2001a.

_____. Acidentes de transporte terrestre em cidade da região sul do Brasil: avaliação da cobertura e qualidade dos dados. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 17, n. 6, p. 1449 - 1456, 2001b.

ANDRADE, S. M. *Acidentes de transporte terrestre em Londrina - PR: análise das vítimas dos acidentes e das fontes de informação*. 1998. 191p. Tese (Doutorado em Saúde Pública) - Faculdade de Saúde Pública. Departamento de Epidemiologia. USP, São Paulo.

BITTENCOURT, S. A.; CAMACHO, L. A. B.; LEAL, M. C. O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 22, n. 1, 2006.

BIGELOW, W.; KARLSON, T.; BEUTEL, P. Using probabilistic linkage to merge multiple data sources for monitoring population health. *Center for Health Systems Research & Analysis. University of Wisconsin*. Madison Junho. 1999, Disponível em: <<http://www.chsra.wisc.edu/CODES/codes2/Probabilistic%20c/Probabilistic%20c.pdf#search=%22data%20linkage%20reliability%22>> Acesso em: 20 ago. 2006.

BOHLAND, A. K.; MELLO JORGE, M. H. P. Mortalidade de menores de um ano de idade na região Sudoeste do Estado de São Paulo. *Revista de Saúde Pública*. São Paulo, v. 33, n. 4, p. 366 - 373, 1999.

BOHLAND, A. K. *Mortalidade de crianças menores de cinco anos, no município de Itararé, São Paulo, em 1993*. 1996. 142p. Dissertação (Mestrado em Saúde Pública) - Faculdade de Saúde Pública. Departamento de Epidemiologia. USP, São Paulo.

BRUM, L.; KUPEK, E. Record linkage and capture-recapture estimates for underreporting of human leptospirosis in a Brazilian health district. *Brazilian Journal Infections Diseases*. Salvador, v. 9, n. 6, p. 515 - 520, 2005.

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

CARVALHO, P. I. *Fatores de risco da mortalidade infantil: análise de coorte de nascidos vivos de mães residentes no Recife no ano de 1999*. 2003. 80p. Dissertação (Mestrado em Saúde Coletiva) - Departamento de Saúde Coletiva. UFPE, Recife.

CAVALCANTE, M. S.; RAMOS JR., A. N.; SILVA, T. M. J. PONTES, L. R. S. K. Transmissão vertical do HIV em Fortaleza: revelando a situação epidemiológica

em uma capital do nordeste. *Revista Brasileira de Ginecologia e Obstetria*. v. 26, n. 2, p. 131 - 138, 2004.

CAVALCANTE, M. S. *Identificação das gestantes infectadas pelo HIV*. 2003, 108p. Dissertação (Mestre em Saúde Pública) - Universidade Federal do Ceará. UFC, Fortaleza.

CAVALCANTE, M. S.; RAMOS JR., A. N.; PONTES, L. R. S. K. Relacionamento de sistemas de informação em saúde: uma estratégia para otimizar vigilância das gestantes infectadas pelo HIV. *Epidemiologia e Serviços de Saúde*. v. 14, n. 2, p. 127 - 133, 2005.

CHRISTEN, P.; CHURCHES, T.; HEGLAND, M. A parallel open source data linkage system. 8th PAKDD'04 (Pacific-Asia Conference on Knowledge Discovery and Data Mining). *Proceedings*. Sydney, 2004. *Springer Lecture Notes in Artificial Intelligence*, (3056). Disponível em: <<http://datamining.anu.edu.au/>>. Acesso em: 21 fev. 2006.

COELI, C. M.; CAMARGO JR., K. R. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. *Revista Brasileira de Epidemiologia*. v. 5, n. 2, p. 185 - 196, 2002.

COELI, C. M.; BLAIS, R.; COSTA, M. C. E.; ALMEIDA, L. M. Relacionamento probabilístico em inquérito domiciliar sobre uso de serviços hospitalares. *Revista de Saúde Pública*. São Paulo, v. 37, n. 1, p. 91 - 99, 2003.

COELI, C. M.; COUTINHO, E. S. F.; VERAS, R. P. O desafio da aplicação da metodologia de captura-recaptura na vigilância do *diabetes mellitus* em idosos: lições de uma experiência no Brasil. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 20, n. 6, p. 1709 - 1720, 2004.

COELI, C. M. *Vigilância do diabetes mellitus em uma população idosa: aplicação da metodologia de captura-recaptura*. 1998. 185p. Tese (Doutorado em Saúde Coletiva) - Instituto de Medicina Social. UERJ, Rio de Janeiro.

COUTINHO, D. P. *Teoria da informação: conceito de entropia e sua aplicação*. Folhas de apoio da disciplina de Compressão e Codificação de Dados (versão provisória 1.0) Seção de Análise de Sinais Departamento de Engenharia de Eletrônica e de Telecomunicações e de Computadores Instituto Superior de Engenharia de Lisboa. 2004. Disponível em: <http://www.deetc.isel.ipl.pt/analisedesinais/ccd/docs/CCD0304FolhasApoio_parte1.pdf>. Acesso em: 21 fev. 2006.

CUNHA, E. M. G. P. Mortalidade infantil por raça/cor. In: BATISTA, L. E.; KALCKMANN, S. *Seminário saúde da população negra do Estado de São Paulo 2004*. São Paulo: Instituto de Saúde, 2005. p. 103 - 115.

_____. *Condicionantes da mortalidade infantil segundo raça/cor no estado de São Paulo, 1997-1998*. 2001. 180p. Tese (Doutorado em Saúde Coletiva) - Departamento de Saúde Coletiva. UNICAMP, Campinas.

FAIR, M. E. Record linkage in an information age society. International Workshop and Exposition on Record Linkage Techniques. *Proceedings*, Chapter 1. Arlington, Virginia, 1997. Disponível em: <www.census.gov/prod/2/gen/96arc/xia-fair.pdf>. Acesso em: 21 fev. 2006

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. *Journal of the American Statistical Association*. v. 40, p. 1183 - 1210, 1969.

FELLEGI, I. P. I. Record linkage and public policy - A dynamic evolution. International Workshop and Exposition on Record Linkage Techniques. *Proceedings*, Chapter 1. Arlington, Virginia, 1997. Disponível em: <<http://www.fscm.gov/working-papers/fellegi.pdf>>. Acesso em: 20 fev. 2006.

FERNANDES, R. M. B. P. *Mortalidade neonatal no município de São Paulo: estudo das causas de morte segundo peso ao nascer, tipo de parto, sexo e idade ao morrer*. 2002. 102p. Dissertação (Mestrado em Saúde Pública) - Faculdade de Saúde Pública. USP, São Paulo.

FERNANDES, D. M. *Concatenamento de informações sobre óbitos e nascimentos: uma experiência metodológica do Distrito Federal - 1989-1991*. 1997. 71p. Tese (Doutorado em Demografia) - Centro de Desenvolvimento e Planejamento Regional. Faculdade de Ciências Econômicas. UFMG, Belo Horizonte.

FERREIRA, V. M. B.; PORTELA, M. C. Avaliação da subnotificação de casos de Aids no município do Rio de Janeiro com base em dados do sistema de Informações Hospitalares do Sistema Único de Saúde. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 15, n. 2, p. 317 - 324, 1999.

FERREIRA, V. M. B. *Análise da subnotificação de casos de Aids no contexto da assistência hospitalar coberta pelo SUS: implicações e intervenções potenciais*. 1999. 85p. Dissertação (Mestrado em Saúde Pública) - Escola Nacional de Saúde Pública. ENSP/Fiocruz, Rio de Janeiro.

FERREIRA, V. M. B.; PORTELA, M. C.; VASCONCELLOS, M. T. L. Fatores associados à subnotificação de pacientes com Aids, no Rio de Janeiro, RJ, 1996. *Revista de Saúde Pública*. São Paulo, v. 34, n. 2, p. 170 - 177, 2000.

GIGLIO, M. R. P.; LAMOUNIER, J. A.; MORAIS NETO, O. L. Via de parto e risco para mortalidade neonatal em Goiânia no ano de 2000. *Revista de Saúde Pública*. São Paulo, v. 39, n. 3, p. 350 - 357, 2005a.

GIGLIO, M. R. P.; LAMOUNIER, J. A.; MORAIS NETO, O. L.; CESAR, C. C. Baixo peso ao nascer em coorte de recém-nascidos em Goiânia-Brasil no ano de 2000. *Revista Brasileira de Ginecologia e Obstetrícia*. v. 27, n. 3, p. 130 - 136, 2005b.

GIGLIO, M. R. P. *Mortalidade neonatal em coorte de recém-nascidos em Goiânia no ano de 2000*. 2003. 101p. Dissertação (Mestrado em Medicina - Pediatria) - Departamento de Pediatria. UFMG, Belo Horizonte.

GOES, S. M. C. *Relacionamento probabilístico entre bases de dados sobre medicamentos e notificação: uma aplicação na vigilância epidemiológica hospitalar da Aids*. 2003. 75p. Dissertação (Mestrado em Saúde Coletiva) - Núcleo de Estudos de Saúde Coletiva. UFRJ, Rio de Janeiro.

GOMES JR., S. C. S.; ALMEIDA, R. T. Identificação de um caso novo de câncer no sistema de informação ambulatorial do Sistema Único de Saúde. *Cadernos Saúde Coletiva*. Rio de Janeiro, v. 12, n.1, p. 57 - 68, 2004.

GOMES, J. O.; SANTO, A. H. Mortalidade infantil em município da região Centro-Oeste Paulista, Brasil, 1990 a 1992. *Revista de Saúde Pública*. São Paulo, v. 31, n. 4, p. 330 - 341, 1997.

GOMES, J. O. *Mortalidade infantil em Presidente Prudente - SP: 1990 a 1992*. 1994. Dissertação (Mestrado em Saúde Pública) - Faculdade de Saúde Pública. Departamento de Epidemiologia. USP, São Paulo.

GU, L.; BAXTER, R.; VICKERS, D.; RAINSFORD, C. *Record linkage: current practice and future directions*. CSIRO Mathematical and Information Sciences, 2003. Disponível em: <http://www.act.cmis.csiro.au/rohanb/PAPERS/record_linkage.pdf>. Acesso em: 21 fev. 2006.

HELENA, E. T. S.; SOUSA, C. A.; SILVA, C. A. Fatores de risco para mortalidade neonatal em Blumenau, Santa Catarina: *linkage* entre bancos de dados. *Revista Brasileira de Saúde Materno Infantil*. v. 5, n. 2, p. 209 - 217, 2005.

LEÃO, B. F.; COSTA, C. G.; FACCHINI, L. A.; BANDARRA, E. B.; GONÇALVES, S. F.; BRETAS JR., N.; FERLA, A. The brazilian health informatics and information policy: building the consensus. *World Congress on Medical Informatics*. San Francisco, California, 2004. Disponível em: <<http://cmbi.bjmu.edu.cn/news/report/2004/medinfo2004/pdf/files/papers/5503Leao.pdf>>. Acesso em: 20 fev. 2006.

LEMONS, K. R. V.; VALENTE, J. G. A declaração de óbito como indicador de sub-registro de casos de Aids. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 17, n. 3, p. 617 - 626, 2001.

LEMOS, K. R. V. *Mortalidade por Aids no estado do Rio de Janeiro e avaliação do sub-registro de casos de Aids tendo como indicador a declaração de óbito*. 1999. 116p. Dissertação (Mestrado em Medicina Social) - Instituto de Medicina Social. UERJ, Rio de Janeiro.

LI, T. T. R.; CONILL, E. M.; KUPEK, E. Excedente de morbimortalidade infantil em áreas homogêneas de risco. *Revista Brasileira de Epidemiologia*. Supl. Especial, p. 213 - 213, 2002.

LI, T. T. R. *Vigilância à saúde e gestão da informação nos SILOS: uma proposta a partir do SINASC no município de São José, SC*. 2000. 139p. Dissertação (Mestrado em Saúde Pública) - Faculdade de Saúde Pública. UFSC, Florianópolis.

MACHADO, C. J.; HILL, K. Determinants of neonatal and post-neonatal mortality in the city of São Paulo. *Revista Brasileira de Epidemiologia*. v. 6, n. 4, p. 345 - 358, 2003.

_____. Relacionamento probabilístico de dados e um procedimento automático para minimizar o problema da incerteza no pareamento de registros. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 20, n. 4, p. 915 - 925, 2004.

_____. Maternal, neonatal and community factors influencing neonatal mortality in Brazil. *Journal Biosocial Science*. v. 37, n. 2, p. 193 - 208, 2005.

MACHADO, C. J. *Early infant morbidity and infant mortality in the city of São Paulo, Brazil: a probabilistic record linkage approach*. 2002. 350p. Tese (Doutorado em Saúde Pública) - Bloomberg School of Public Health. Johns Hopkins University, Baltimore.

_____. A literature review of record linkage procedures focusing on infant health outcomes. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 20, n. 2, p. 362 - 371, 2004.

MARTINS, C. B. G.; ANDRADE, S. M. Epidemiologia dos acidentes e violências entre menores de 15 anos em município da região sul do Brasil. *Revista Latino-Americana de Enfermagem*. v. 13, n. 4, p. 530 - 537, 2005a.

_____. Causas externas entre menores de 15 anos em cidade do sul do Brasil: atendimentos em pronto-socorro, internações e óbitos. *Revista Brasileira de Epidemiologia*. v. 8, n. 2, p. 194 - 204, 2005b.

MARTINS, C. B. G. *Epidemiologia dos acidentes e violências em menores de 15 anos em Londrina, Paraná*. 2004. 155p. Dissertação (Mestrado em Saúde Coletiva) – Centro de Estudos Superiores. Universidade Estadual de Londrina, Londrina.

MARTINS, E. F.; VELASQUEZ-MELENDZ, G. Determinantes da mortalidade neonatal a partir de uma coorte de nascidos vivos, Montes Claros, Minas Gerais, 1997-1999. *Revista Brasileira de Saúde Materno Infantil*. v. 4, n. 4, p. 405 - 412, 2004.

MARTINS, E. F. *Mortalidade neonatal em Montes Claros-Minas Gerais, no período de 1997 a 1999*. 2002. 136p. Dissertação (Mestrado em Enfermagem) – Escola de Enfermagem. UFMG, Belo Horizonte.

MASON, C. A. *Probabilistic Linkage: Issues and Strategies*. National Early Hearing Detection and Intervention Conference. Washington D.C., 2006. Disponível em: <www.infanthearing.org/meeting/ehdi2006/presentations/MasonC_EHDI2006.ppt>. Acesso em: 20 ago. 2006.

MELLO, A. L. L. C. *Prevalência e sobrevivência de casos de acidente vascular encefálico no município do Rio de Janeiro no ano de 1998*. 2003. 90p. Dissertação (Mestrado em Saúde Pública) - Escola Nacional de Saúde Pública. ENSP/Fiocruz, Rio de Janeiro.

MORAIS NETO, O. L.; BARROS, M. B. A. Fatores de risco para mortalidade neonatal e pós-neonatal na região Centro-Oeste do Brasil: *linkage* entre bancos de dados de nascidos vivos e óbitos infantis. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 16, n. 2, p. 477 – 485, 2000.

MORAIS NETO, O. L. *A mortalidade infantil no município de Goiânia: uso vinculado do SIM e SINASC*. 1996. 129p. Dissertação (Mestrado em Saúde Coletiva) - Universidade Estadual de Campinas. UNICAMP, Campinas.

_____. *Padrão espacial da mortalidade infantil em Goiânia-Goiás, nas coortes de nascidos vivos 11 entre 1992 a 1996*. 2002, 150 p. Tese (Doutorado em Saúde Coletiva) - Universidade Estadual de Campinas. UNICAMP, Campinas.

NORONHA, C. P.; SILVA, R. L.; THEME FILHA, M. M. Concordância de dados das declarações de óbitos e de nascidos vivos para a mortalidade neonatal no município do Rio de Janeiro. *IESUS*. v. 4, p. 57 - 65, 1997.

ORTIZ FLORES, L. P. *Características da mortalidade neonatal no Estado de São Paulo*. 1999. 183 p. Tese (Doutorado em Saúde Pública) - Faculdade de Saúde Pública. Departamento de Epidemiologia. USP, São Paulo.

PACHECO, R. C.; KERN, V. Uma ontologia comum para a integração de bases de informações e conhecimento sobre ciência e tecnologia. *Ciência da Informação*. Brasília, DF, v. 30, n. 3, 2001. Disponível em: <<http://www.ibict.br/cionline/viewarticle.php?id=231>>. Acesso em: 20 fev. 2006.

PEPE, V. L. E. *Internações e reinternações psiquiátricas no Estado do Rio de Janeiro*. 2002. 202p. Tese (Doutorado em Medicina Preventiva) – Departamento de Medicina Preventiva. USP, São Paulo.

PETERS, L. R. *As mães adolescentes e os filhos delas: veredas: nascimento e morte no primeiro ano de vida*, Joinville, SC, 1995. 2001. 105 p. Dissertação (Mestrado em Saúde Pública) - Faculdade de Saúde Pública, Departamento de Saúde Materno-Infantil. USP, São Paulo.

PINHEIRO, A. M. C. M. *Avaliação dos sistemas de informação de nascidos vivos e de mortalidade para a obtenção da mortalidade neonatal em Ilhéus, Bahia*. 2003. 108 p. Tese (Doutorado em Saúde Pública) - Faculdade de Saúde Pública, Departamento de Epidemiologia. USP, São Paulo.

PINHO, S. F. C. *Avaliação da qualidade de dados pela não-conformidade*. 2001 120p. Dissertação (Mestrado em Engenharia de Produção) - Programa de Engenharia de Sistemas e Computação da Coordenação dos Programas de Pós-graduação de Engenharia. COPPE/UFRJ, Rio de Janeiro.

PITTA, A. M. R.; MAGAJEWSKI, F. R. L. Políticas nacionais de comunicação em tempos de convergência tecnológica: uma aproximação ao caso da saúde. *Interface - Comunicação, Saúde, Educação*. v. 4 , n. 7, p. 61 - 70, 2000.

PORTELA, M. C.; SCHRAMM, J. M. A.; PEPE, V. L. E.; NORONHA, M. F.; PINHO, C. A. M.; CIANELI, M. P. Algoritmo para a composição de dados por internação a partir do sistema de informações hospitalares do sistema único de saúde (SIH/SUS) - Composição de dados por internação a partir do SIH/SUS. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 13, n. 4, p. 771 - 774, 1997.

ROSSO, C. F. W. *Mortalidade neonatal em Goiânia: fatores prognósticos de recém-nascidos de risco*. 2003. 60p. Dissertação (Mestrado em Medicina Tropical) - Universidade Federal de Goiás, Goiânia.

SARACENI, V.; LEAL, M. C. Avaliação da efetividade das campanhas para eliminação da sífilis congênita na redução da morbi-mortalidade perinatal: município do Rio de Janeiro, 1999-2000. *Cadernos de Saúde Pública*. Rio de Janeiro, v. 19, n. 5, p. 1341 - 1349, 2003.

SARACENI, V.; CRUZ, M. M.; LAURIA, L. M. DUROVNI, B. Trends and characteristics of AIDS mortality in the Rio de Janeiro city after the introduction of highly active antiretroviral therapy. *Brazilian Journal of Infection Disease*. v. 9, n. 3, p. 209 - 215, 2005.

SARACENI, V. *Avaliação da efetividade das campanhas para eliminação da sífilis congênita do município do Rio de Janeiro, 1999-2000*. 2005. 111p. Tese (Doutorado em Saúde Pública) - Escola Nacional de Saúde Pública Sérgio Arouca. ENSP/Fiocruz, Rio de Janeiro.

- SARINHO, S. W. *Mortalidade neonatal na cidade do Recife: um estudo caso-controle*. 1998. 136p. Tese (Doutorado em Medicina - Pediatria) - Centro de Ciências da Saúde. UFPE, Recife.
- SARINHO, S. W.; MELO FILHO, D. A.; SILVA, G. A. P.; LIMA, M. C. Fatores de risco para óbitos neonatais no Recife: um estudo de caso-controle. *Jornal de Pediatria*. Rio de Janeiro, v. 77, n. 4, p. 294 - 298, 2001.
- SELIG, L.; BELO, M.; CUNHA, A. J. L. A.; TEIXEIRA, E. G.; BRITO, R.; LUNA, A. L.; TRAJMAN, A. Óbitos atribuídos à tuberculose no Estado do Rio de Janeiro. *Jornal Brasileiro de Pneumologia*. v. 30, n. 4, p. 335 - 342, 2004.
- SILVA, A. B. O. O sistema de informações estatísticas no Brasil e as relações entre seus produtores e usuários. *Ciência da Informação*. Brasília, DF, v. 34, n. 2, p. 62 - 69, 2005. Disponível em: <www.ibict.br/cienciadainformacao/include/getdoc.php?id=1558&article=687&mode=pdf>. Acesso em: 22 abr. 2006.
- SILVA, J. P. L. *A desinstitucionalização e o processo de reformulação da assistência psiquiátrica no Rio de Janeiro no período 1995-2000*. 2003. 100 p. Dissertação (Mestrado em Saúde Pública) - Escola Nacional de Saúde Pública. ENSP/Fiocruz, Rio de Janeiro.
- SOARES, E. P. *Associação entre peso ao nascer e mortalidade infantil no município de Campos dos Goytacazes*. 2003, 140p. Dissertação (Mestrado em Saúde Coletiva) – Núcleo de Estudos de Saúde Coletiva. UFRJ, Rio de Janeiro.
- VIEIRA, R. A. *Sobrevida após fratura proximal de fêmur em pacientes idosos atendidos pelo SUS no município do Rio de Janeiro*. 2004. 75p. Dissertação (Mestrado em Saúde Coletiva) - Núcleo de Estudos de Saúde Coletiva. UFRJ, Rio de Janeiro.
- WAND, Y.; WANG, R. Y. Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM - Association for Computing Machinery*. New York, NY, v. 39, n. 11, p. 86 - 95, 1996. Disponível em: <http://web.mit.edu/tdqm/www/tdqmpub/WandWangCACMNov96.pdf> Acesso em: 14 Jun. 2006.
- WEIRICH, C. F.; ANDRADE, A. L. S. S.; TURCHI, M. D.; SILVA, S. A.; MORAIS-NETO, O. L.; MINAMISAVA, R.; MARQUES, S. M. Mortalidade neonatal em unidades de cuidados intensivos no Brasil Central. *Revista de Saúde Pública*. São Paulo, v. 39, n. 5, p. 775 - 781, 2005.
- WINKLER, W. E. *Machine learning, information retrieval, and record linkage*. U.S. Bureau of the Census Statistical Research. Washington, 2003. Disponível em: <<http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf>>. Acesso em: 21 fev. 2006.

_____. *Record linkage software and methods for merging administrative lists*. U.S. Bureau of the Census Methodology and Standards Directorate Statistical Research Division, Statistical Research Report Series. 2001. Washington D.C. Disponível em: <<http://www.census.gov/srd/papers/pdf/rr2001-03.pdf>>. Acesso em: 20 fev. 2006.