Conquistas e desafios para o emprego das técnicas de *record linkage* na pesquisa e avaliação em saúde no Brasil*

doi: 10.5123/S1679-49742015000400023

Achievements and challenges for employing record linkage techniques in health research and evaluation in Brazil

Cláudia Medina Coeli[1]
Rejane Sobrino Pinheiro[1]
Kenneth Rochel de Camargo Jr.[2]

[1]Universidade Federal do Rio de Janeiro, Instituto de Estudos em Saúde Coletiva, Rio de janeiro-RJ, Brasil
[2]Universidade do Estado do Rio de Janeiro, Instituto de Medicina Social, Rio de Janeiro-RJ, Brasil

Resumo

Objetivo: A disponibilidade em formato digital de bases epidemiológicas, administrativas e clínicas, assim como de técnicas de record linkage (RL) que permitem sua vinculação, consolidaram, nas últimas décadas, o uso de dados secundários na pesquisa e avaliação em saúde. Neste artigo, são discutidas as técnicas de RL, seus aspectos metodológicos e éticos e sua aplicação para a construção de registros populacionais. Por fim, reflete-se sobre os desafios para a pesquisa baseada no uso de técnicas de RL no Brasil, em função da adoção de um novo marco legal para a proteção de dados pessoais. Entre as conclusões, ressalta-se a necessidade de formular o quadro de referência legal e operacional para as atividades de vinculação de bases de dados em nosso país, quer para a pesquisa, quer para a gestão.

Palavras-chave: Registro Médico Coordenado; Bases de Dados como Assunto; Privacidade.

Abstract

Objective*: The availability of epidemiological, clinical and administrative databases in digital format, in addition to the development of record linkage techniques (RL) which enable researchers to link them together, have helped to consolidate the use of secondary data in health research and evaluation in recent decades. In this article we present a discussion of RL techniques, addressing methodological and ethical aspects as well as their application in building population records. Finally, we consider the challenges for research based on the use of RL techniques in Brazil, due to the adoption of a new legal framework for personal data protection. In conclusion, we emphasize the need to develop a legal and operational framework as a foundation for activities involving record linkage in our country, whether it be for research or managerial purposes.*

Key words*: Medical Record Linkage, Databases as Topic; Privacy.*

* Os autores recebem bolsa de pesquisa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)/Ministério da Ciência, Tecnologia e Inovação (MCTI) – Cláudia Medina Coeli, Processo no 304101/2011-7; Rejane Sobrino Pinheiro, Processo no 309728/2012-6; e Kenneth Rochel de Camardo Jr., Processo no 300686/2013-7 – e da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) – Cláudia Medina Coeli, Processo no E26/102.771/2012; e Kenneth Rochel de Camardo Jr., Processo no E-26/102.900/2012.

Endereço para correspondência:
Cláudia Medina Coeli – Universidade Federal do Rio de Janeiro, Instituto de Estudos em Saúde Coletiva, Av. Horácio Macedo, s/n (Próximo a Prefeitura Universitária da UFRJ), Cidade Universitária, Ilha do Fundão, Rio de Janeiro-RJ, Brasil. CEP: 21941-598
E-mail: coeli@iesc.ufrj.br

Introdução

Nos últimos anos, ocorreu uma expansão sem precedentes do acervo de dados disponíveis para utilização na pesquisa, vigilância e avaliação em saúde. Esse crescimento é resultado do aumento do uso da internet no acesso a informações, participação em redes sociais e utilização de aplicativos web, especialmente desde dispositivos móveis. Termos como *big data*, mineração de dados (*data mining*), de texto (*text mining*) e da *web* (*web mining*) passaram a fazer parte do vocabulário da pesquisa em saúde populacional.[1]

> *A disponibilidade em formato digital de bases epidemiológicas, administrativas e clínicas, assim como de técnicas de record linkage que permitem sua vinculação, consolidaram, nas últimas décadas, o uso de dados secundários na pesquisa e avaliação em saúde.*

É indiscutível que o maior volume e variedade de dados, juntamente com a aplicação de técnicas para extração de conhecimento, contribui para a pesquisa e avaliação em saúde. Entretanto, alguns problemas podem limitar os ganhos potenciais do uso desse rico acervo de dados. O perfil selecionado de usuários dos serviços disponíveis na internet, a aplicação pelos gestores desses serviços de filtros que não são divulgados e a vulnerabilidade para a identificação de correlações entre eventos explicados somente pelo acaso são questões que devem ser consideradas.[2,3] Iniciativas como, por exemplo, o programa Big Data to Knowledge (BD2K), implementado pelo National Institute of Health dos Estados Unidos da América (NIH/USA), buscam explorar o potencial do *big data* na pesquisa em saúde populacional.[4]

Ainda que a era do *big data* em saúde se encontre em seu estágio inicial,[5] a disponibilidade em formato digital de bases epidemiológicas, administrativas e clínicas, assim como de técnicas de *record linkage* que permitem sua vinculação, consolidaram, nas últimas décadas, o uso de dados secundários na pesquisa e avaliação em saúde.

Este artigo tem por objetivo discutir as técnicas de *record linkage*, seus aspectos metodológicos e éticos, assim como sua aplicação na construção de registros populacionais. Por fim, é feita uma reflexão sobre os desafios para a pesquisa baseada no uso de técnicas de RL no Brasil.

Record linkage

Record linkage (RL) é o processo de (i) combinação de registros de um mesmo indivíduo presentes em duas bases de dados distintas, ou de (ii) identificação, em uma mesma base, de registros que se referem ao mesmo indivíduo.[6]

A combinação de bases de dados traz como vantagem permitir que hipóteses etiológicas possam ser melhor exploradas com o uso de dados secundários, uma vez que torna disponível para análise um conjunto maior de variáveis, facilitando o ajuste para variáveis de confusão. Outra vantagem da aplicação do RL é possibilitar a realização de estudos longitudinais, retrospectivos ou prospectivos.[7] Pode-se, inclusive, acompanhar a ocorrência de eventos de saúde no curso de vida de um indivíduo, desde o nascimento até a morte.[8] O RL também é empregado na melhora da qualidade de dados de bases de saúde, permitindo a identificação e eliminação de registros duplicados, o preenchimento de dados faltantes, a correção de campos registrados com valores não válidos, assim como a correção de subregistro.[9]

Técnicas de RL

Quando as bases a serem vinculadas apresentam um identificador unívoco, o RL é relativamente simples, pois a maioria dos programas gerenciadores de bases de dados e dos pacotes estatísticos trazem comandos para a realização da vinculação baseada em um campo-chave. Entretanto, a maioria das bases de interesse para estudos na área de saúde não apresenta um identificador unívoco comum, sendo necessário empregar múltiplos campos que armazenam dados pessoais. Nesse caso, o RL envolve várias etapas, empregando distintas estratégias em cada uma delas. Uma discussão abrangente e atualizada sobre as diferentes estratégias disponíveis pode ser encontrada no livro de Peter Christen.[6] São apresentadas a seguir, de forma resumida, as principais etapas envolvidas em RL.

A primeira etapa consiste no pré-processamento das bases, que abrange a limpeza de dados, a padronização de conteúdos e formatos, a quebra dos campos 'nome' e 'endereço' em seus componentes (*parsing*) e a criação de códigos fonéticos. O pré-processamento é realizado apenas uma vez em cada base, e as bases limpas e padronizadas resultantes podem ser utilizadas em projetos futuros de vinculação de dados.

A etapa seguinte vem a ser a indexação ou blocagem (do termo em inglês *blocking*), com o propósito de reduzir o número de *links* de registros que serão enviados para comparação e classificação. O processo tradicional consiste em particionar os dois arquivos, relativamente aos valores de um ou mais campos constitutivos da chave de indexação – por exemplo: sexo + código fonético do primeiro nome + código fonético do último nome. As comparações de registros são, então, restritas a registros que concordam no valor dessa chave. A indexação é um *trade-off* entre comparar muitos pares de registros e a vulnerabilidade à perda de *links* verdadeiros. Com o objetivo de minimizar essas perdas, os projetos de vinculação de bases são realizados em múltiplos passos, empregando-se, em cada um deles, uma chave de indexação formada por uma combinação diferente de campos.

Os *links* formados na etapa anterior são processados nas etapas de comparação e classificação. Na comparação, são empregadas funções exatas ou aproximadas, que indicam semelhança entre dois atributos (concordância parcial). As funções de comparação são específicas para cada tipo de dado – por exemplo: cadeias de caracteres, números, datas. Para cada *link* de registros, o resultado relativo à similaridade de cada atributo é armazenado em um vetor numérico. Finalmente, o vetor de similaridade é empregado no processo de classificação automática, considerando-se que se a semelhança entre dois registros é tão evidente, o mais provável é ambos os registros referirem-se ao mesmo indivíduo. O resultado final é a classificação do *link* em par verdadeiro ou falso; senão, em caso de dúvida sobre o par, os registros são encaminhados para revisão manual.

As técnicas de *linkage* mais frequentemente empregadas são a determinística e a probabilística. As técnicas determinísticas usam funções exatas de comparação e abordagem classificatória fundada em regras definidas e baseadas em conhecimento especializado, sendo necessário desenvolver rotinas de computação específicas para cada projeto. Já a técnica probabilística, baseada no modelo de Fellegi & Sunter,[10] utiliza funções de comparação aproximadas. Pesos diferentes são atribuídos a cada campo com base em seu poder de discriminação e vulnerabilidade ao erro, sendo calculado um escore indicativo do quão verossímil é dois registros pertencerem ao mesmo indivíduo. Trata-se de uma solução geral, aplicável a diferentes projetos de *linkage*. Uma série de *softwares open source* e comerciais implementam essa técnica,[6] que não requer habilidades especiais em programação. O *software* OpenReclink (http://reclink.sourceforge.net/),[11] desenvolvido pelos autores deste artigo, implementa o modelo de Fellegi & Sunter e representa uma evolução do *software* Reclink.[12] A versão *open*, além do código aberto, traz como novidades seu caráter de multiplataforma, multilinguagem e uma revisão da rotina de deduplicação. A versão Linux já está em uso corrente, no laboratório do Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro (IESC-UFRJ) (LinkDataPop), no desenvolvimento de projetos de grande porte, apresentando boa performance e estabilidade. Uma alternativa aos modelos anteriormente descritos é o uso de rotinas baseadas em aprendizado de máquina (*machine learning*). Esse modelo é implementado pelo *software* ChoiceMaker (http://oscmt.sourceforge.net/), o mesmo adotado pelo The Centre for Health Record Linkage (CheReL) (NSW, Austrália).[13] O modelo de classificação baseia-se na utilização de regras, que são ponderadas aplicando-se técnicas de aprendizado de máquina. O peso de cada regra é empregado para o cálculo de um escore que indica quão verossímil é dois registros pertencerem ao mesmo indivíduo.

A figura 1 ilustra a tela inicial do programa OpenReclink.

Independentemente do modelo adotado, uma parte dos *links* classificados como objeto de dúvida na etapa de classificação automática necessitará ser submetida a revisão manual, para sua classificação definitiva. Esta etapa demanda, portanto, o maior consumo de recursos humanos e horas para sua realização. Por exemplo, em um estudo realizado por estes pesquisadores, com o objetivo de identificar o números de filhos vivos anteriores de mulheres que tiveram filhos nascidos vivos no estado do Rio de Janeiro entre 2007 e 2008, foi processada uma base com cerca de 2.400.000 registros, e consumidas cerca de 800 horas no processo de revisão. Para otimizar o processo de revisão manual, tomando-se por base a experiência de especialistas em revisão, foi desenvolvida uma rotina de classificação automática de *links* baseada em 20 critérios distintos, gerando uma nota final para cada *link*. O algoritmo foi testado em uma aplicação envolvendo o *linkage* entre as bases do Sistema de Informação de Agravos de Notificação – Aids (Sinan-Aids) e do Sistema de Informações sobre Mortalidade (SIM), demonstrando boa performance.[14]

Figura 1 – Tela inicial do programa OpenReclink, com menu de processos no topo

A última etapa de um projeto de *linkage* de bases de dados deve envolver a avaliação da qualidade do processo em termos de proporção de pares que deixaram de ser formados, a despeito de se referirem ao mesmo indivíduo, e de pares falsos, i.e., associando indivíduos diferentes. Todavia, essa etapa é pouco valorizada em estudos brasileiros que empregam técnicas de RL.[15] É preciso dedicar maior atenção à avaliação dos processos de *linkage* realizados no Brasil, haja vista a informação sobre endereço registrada em nossas bases de dados nem sempre permitir sua utilização no processamento automático, e alguns sobrenomes (ex. Silva) e nomes (ex. Maria) serem bastante frequentes no Brasil,[16] aumentando a vulnerabilidade para a formação de pares falso-positivos na etapa de classificação automática.

Ética e privacidade

Desde a publicação dos artigos seminais sobre RL de Newcombe e colaboradores[17] e de Fellegi & Sunter[18] ao final das décadas de 1950 e 1960, respectivamente, novas tecnologias de informação foram introduzidas, levando a um aumento expressivo da capacidade de armazenamento e processamento de bases de dados. O livro de Peter Christen,[6] já mencionado neste artigo, discute várias soluções capazes de otimizar o RL, tanto em termos de tempo de processamento como de acurácia. Hoje, a maior barreira para a aplicação de técnicas de RL na pesquisa e avaliação em saúde não reside na técnica e sim na ética, uma vez que a maioria dos processos de RL demanda o acesso a dados pessoais.

Vários países do mundo contam com legislações orientadas à proteção individual, no que tange à utilização de dados pessoais em pesquisas e avaliações de saúde. Essas legislações determinam que o acesso aos dados somente possa ser feito mediante autorização do indivíduo. Reconhecendo a necessidade da busca de um equilíbrio entre a garantia dos direitos individuais e os potenciais ganhos para a saúde da coletividade, essas mesmas legislações, geralmente, tratam o uso para pesquisa como uma exceção legítima, permitindo que

o acesso a dados pessoais seja feito sem esse consentimento, desde que o projeto de pesquisa atenda a uma série de requisitos. Entre esses requisitos, a proposta do estudo deve estar de acordo com certas diretrizes éticas para ser aprovada por um comitê de ética em pesquisa habilitado. As diretrizes éticas para estudos epidemiológicos adotadas pelo Council for International Organizations of Medical Sciences (CIOMS)[19] preveem a liberação da exigência do consentimento livre e esclarecido quando: a pergunta da pesquisa é relevante para a sociedade; a pergunta implica risco mínimo para o sujeito da pesquisa; os pesquisadores asseguram a não violação de direitos individuais e coletivos; são adotadas normas que garantem a segurança, a privacidade e o anonimato das informações; e a solicitação de consentimento inviabilizaria o estudo, pela dificuldade em operacionalizá-lo.

Em uma revisão sistemática, encontrou-se uma proporção elevada de pedidos de consentimento para RL na maioria dos estudos observados,[20] embora quase todos os casos em que o consentimento foi solicitado envolvessem coleta de dados primários. A experiência com a solicitação de consentimento diretamente à população geral – por exemplo, em programas de doação de órgãos – sugere que sua operacionalização seja mais problemática ainda, sofrendo, inclusive, influência do tipo de questão apresentada pela solicitação – consentimento explícito (*opt in*) *versus* consentimento implícito (*opt out*).[21] Um relatório elaborado pelo Committee on Health Research and the Privacy of Health Information/Institute of Medicine[22] recomenda que a solicitação do consentimento esclarecido sobre o uso de informações pessoais para fins da pesquisa e avaliação em saúde seja substituída por uma maior ênfase na implementação de medidas que garantam transparência ao uso, segurança e confidencialidade das informações utilizadas. Austrália,[23] Canadá[24] e País de Gales,[25] por exemplo, têm experiência consolidada na implantação de unidades de vinculação de dados, que proporcionam um ambiente seguro para o armazenamento e processamento de dados identificados, e ao mesmo tempo, adotam modelos de governança de dados que asseguram um equilíbrio entre a preservação do direito à privacidade e os ganhos potenciais advindos de seu uso na avaliação e pesquisa populacional.

No Brasil, em 2011, foi promulgada a Lei de Acesso à Informação (Lei ordinária no 12.527, de 18 de novembro de 2011),[26] que regulamenta o direito ao acesso à informação pública. No que diz respeito às informações pessoais, a Lei determina que esse acesso seja feito exclusivamente com o consentimento da pessoa. A mesma Lei admite, entretanto, a desobrigação desse consentimento em algumas situações, incluindo a realização de pesquisa científica. Em julho de 2015, foi encerrado o período de consulta pública ao anteprojeto de lei para a proteção de dados pessoais (http://pensando.mj.gov.br/dadospessoais/). O texto desse anteprojeto também prevê que a pesquisa científica seja tratada como exceção no que tange à exigência do consentimento, sem apresentar maiores detalhes – a serem definidos, provavelmente, mediante regulamentação complementar. Já a Resolução do Conselho Nacional de Saúde (CNS) nº 466, de 12 de dezembro de 2012, ao regulamentar a pesquisa envolvendo seres humanos (http://conselho.saude.gov.br/resolucoes/2012/Reso466.pdf), prevê a dispensa do consentimento em seu capítulo IV, item 8:

> [...] Nos casos em que seja inviável a obtenção do Termo de Consentimento Livre e Esclarecido ou que esta obtenção signifique riscos substanciais à privacidade e confidencialidade dos dados do participante ou aos vínculos de confiança entre pesquisador e pesquisado, a dispensa do TCLE deve ser justificadamente solicitada pelo pesquisador responsável ao Sistema CEP/CONEP, para apreciação, sem prejuízo do posterior processo de esclarecimento.

Portanto, de forma semelhante a outros países, a legislação brasileira, ao mesmo tempo que busca assegurar o direito à privacidade, admite a liberação do pedido de consentimento de indivíduos para utilização de suas informações em situações de pesquisa, buscando alcançar um equilíbrio entre a necessidade de proteção dos direitos individuais e os potencias ganhos para a sociedade.

Unidades de vinculação de dados (*data linkage units*)

Com relação às aplicações de RL, uma das mais interessantes é sua utilidade para a construção de registros de saúde. Esses registros são formados pela integração rotineira de bases de saúde (vitais, administrativas, clínicas...) e de outros setores (educação, por ex.). Uma vez constituídos, os registros permitem, de forma

eficiente e preservando a privacidade, gerar bases específicas capazes de responder a questões de interesse para a pesquisa e a avaliação em saúde. Ademais, o RL permite que dados coletados primariamente possam ser integrados com as bases do registro, auxiliando o seguimento longitudinal de participantes. Os registros são implantados em unidades dotadas de infraestrutura suficiente para garantir o processamento e a custódia segura de grande volume de dados identificados. Lembre-se que todas as unidades realizam a governança dos dados, segundo protocolos publicizados.

Os países com experiências mais consolidadas na implantação desse modelo de unidade de vinculação de dados são Austrália, Canadá, País de Gales e Escócia. Desde 2007, a Austrália vem implementando uma rede nacional de unidades de *linkage* de dados (The Population Health Research Network [PHRN]) (http://www.phrn.org.au/about-us/overview). Mais recentemente, foi implantada no Reino Unido uma rede formada por quatro unidades de *linkage* de dados, The Farr Institute of Health Informatics Research (http://www.farrinstitute.org/). Embora os modelos dessas unidades variem em alguns aspectos,[27] todos têm como missão (i) integrar bases de dados de natureza diversa e (ii) disponibilizar bases derivadas não identificadas para fins de avaliação e pesquisa. Essas unidade também são encarregadas de desenvolver técnicas de processamento e análise de dados secundários, além de promover a capacitação de técnicos em saúde e pesquisadores nas temáticas de RL e análise de dados secundários.

Em 2008, foi criada a International Health Data Linkage Network (IHDLN) (http://www.ipdln.org/), uma rede internacional com o propósito de estimular a cooperação entre pesquisadores e unidades de *data linkage* em temas como tecnologia de informação, RL e análise de dados secundários. Em 2014, no sentido de melhor caracterizar o tipo de pesquisa realizada pelos membros dessa rede internacional, ela passou a se denominar The International Population Data Linkage Network (IPDLN). Atualmente, são afiliadas à rede 23 unidades *linkage* de dados, localizadas na Austrália (nove), Canadá (sete), Reino Unido (cinco), Alemanha (uma) e Nova Zelândia (uma). Considerando-se o país de origem dos pesquisadores membros – 34 países –, a representatividade geográfica é maior. O Brasil, até o momento da elaboração deste artigo, é o único país da América do Sul com membros participantes.

Experiência brasileira em RL

Artigos sobre RL em saúde no Brasil começaram a ser publicados em meados da década de 1990.[28] Desde então, essa produção vem aumentando, em número e variedade de autores e instituições de origem. Uma busca realizada na base PubMed em 11 de outubro de 2015, empregando a chave 'record linkage AND (Brasil OR Brazil)', levantou 68 artigos. Contudo, é importante considerar que esse número não reflete a produção total de manuscritos, muitos publicados em periódicos não indexados na base PubMed. Um exemplo destes é o número temático do periódico *Cadernos de Saúde Coletiva* dedicado ao RL (http://www.cadernos.iesc.ufrj.br/cadernos/index.php/features-sp-417739839/2006/no2-abr-jun).

Concomitantemente às aplicações de RL na pesquisa, também houve interesse de gestores na incorporação do RL para reduzir a subnotificação de eventos e melhorar a qualidade das bases de dados em saúde. A parceria entre pesquisadores e gestores dos níveis federal, estadual e municipal aconteceu via projetos voltados ao desenvolvimento de soluções de RL, seja no apoio à vigilância e monitoramento em saúde, seja na capacitação de técnicos de saúde sobre o tema. Resultado positivo dessa parceria foi a adoção, a partir de 2004, de técnicas de RL para correção da subnotificação de casos de aids no Sinan.[29]

Frente às inúmeras possibilidades da utilização do RL e seu potencial de contribuição para o aprimoramento da qualidade da informação em saúde, e aos recentes avanços no marco legal e regulatório (Lei de Acesso à Informação e Resolução CNS nº 466/2012), mostram-se necessárias mudanças na forma como a pesquisa em RL vem sendo realizada no país. É mister que o Brasil adote um modelo nacional de vinculação de bases de dados populacionais, e a experiência internacional, baseada em unidades de vinculação de dados, pode servir como uma referência para a construção do modelo brasileiro. É fundamental que questões como a especificidade da legislação do país, nossa experiência na produção e disseminação de dados secundários, além do grande volume e abrangência nacional das bases brasileiras sejam considerados na formulação do modelo a ser adotado. Também é importante criar mecanismos que permitam a pesquisadores, gestores, custodiantes de dados e representantes da sociedade civil trabalharem juntos, na adaptação do modelo internacional às espe-

cificidades brasileiras. Uma iniciativa nesse sentido foi o 'Seminário sobre Vinculação de Bases de Dados na Saúde', realizado no Rio de Janeiro-RJ, em outubro de 2014. Organizado pelo Departamento de Informática do Sistema Único de Saúde (Datasus)/Ministério da Saúde, o seminário contou com a participação de atores envolvidos no processo de geração, custódia e análise de dados no âmbito da saúde, além de representantes convidados de três países com marcante desenvolvimento nessa área: Canadá, País de Gales e Austrália. É importante que essa iniciativa tenha seguimento, na formulação de um quadro de referência legal e operacional para as atividades de vinculação de bases de dados em nosso país, quer para a pesquisa e análise, quer para a gestão.

A definição pelo país de seu próprio modelo institucional para uma questão tão complexa demandará algum tempo. Entretanto, a simples manutenção do *status quo* é indesejável, tendo em vista que algumas práticas, de uso corrente até aqui, podem conflitar com uma nova legislação sobre acesso a dados. Por isso, é fundamental que se estabeleça, o mais brevemente possível, um conjunto de regras de transição para o acesso a bases de dados identificadas que sejam compatíveis com a legislação vigente, sem prejuízo à continuidade dos projetos em andamento, até que se disponha de um novo modelo, capaz de garantir os avanços e ganhos conquistados com a pesquisa baseada em RL no Brasil.

Referências

1. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Comput Biol. 2012 Jul;8(7):e1002616.
2. Ruths D, Pfeffer J. Social media for large studies of behavior. Science. 2014 Nov;346(6213):1063–4.
3. Khoury MJ, Ioannidis JP. Medicine. Big data meets public health. Science. 2014 Nov;346(6213):1054–5.
4. Kaplan RM, Riley WT, Mabry PL. News from the NIH: leveraging big data in the behavioral sciences. Transl Behav Med. 2014 Sep;4(3):229–31.
5. Chiavegatto Filho ADP. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol Serv Saude. 2015 abr-jun;24(2):325–32.
6. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. New York: Springer-Verlag Berlin Heidelberg; 2012.
7. Vidal EI, Coeli CM, Pinheiro RS, Camargo Júnior KR. Mortality within 1 year after hip fracture surgical repair in the elderly according to postoperative period: a probabilistic record linkage study in Brazil. Osteoporos Int. 2006 Oct;17(10):1569–76.
8. Kramer MR, Dunlop AL, Hogue CJ. Measuring women's cumulative neighborhood deprivation exposure using longitudinally linked vital records: a method for life course MCH research. Matern Child Health J. 2014 Feb;18(2):478–87.
9. Bartholomay P, Oliveira GP, Pinheiro RS, Vasconcelos AMN. Melhoria da qualidade das informações sobre tuberculose a partir do relacionamento entre bases de dados. Cad Saude Publica.2014 nov;30(11):2459–70.
10. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques. New York: Springer-Verlag; 2007.
11. Camargo Júnior KR, Coeli CM. Going open source: some lessons learned from the development of OpenRecLink. Cad Saude Publica. 2015 Feb;31(2):257–63.
12. Camargo Júnior KR , Coeli CM. Reclink: an application for database linkage implementing the probabilistic record linkage method. Cad Saude Publica. 2000 Apr-Jun;16(2):439–47.
13. Taylor LK, Irvine K, Iannotti R, Harchak T, Lim K. Optimal strategy for linkage of datasets containing a statistical linkage key and datasets with full personal identifiers. BMC Med Inform Decis Mak. 2014 Sep;14:85.
14. Lucena F. Redes neurais artificiais na identificação de registros de um mesmo indivíduo no relacionamento de bases de dados distintas [dissertação]. Rio de Janeiro (RJ): Universidade Federal do Rio de Janeiro; 2013.
15. Coeli CM. A qualidade do linkage de dados precisa de mais atenção. Cad Saude Publica. 2015 jul;31(7):1349–50.
16. Coeli CM, Camargo Júnior KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol. 2002;5(2):185–96.
17. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records: computers can be used to extract "follow-up" statistics of families from files of routine records. Science. 1959 Oct;130(3381):954–9.

18. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc. 1969 Dec;64(328):1183–210.

19. Council for International Organizations of Medical Sciences. International ethical guidelines on epidemiological studies. Geneva: Council for International Organizations of Medical Sciences; 2009.

20. Silva ME, Coeli CM, Ventura M, Palacios M, Magnanini MM, Camargo TM, et al. Informed consent for record linkage: a systematic review. J Med Ethics. 2012 Oct;38(10):639–42.

21. Johnson EJ, Goldstein DG. Defaults and donation decisions. Transplantation. 2004 Dec;78(12):1713–6.

22. Sharyl JN, Laura AL,Lawrence OG, editors; Committee on Health Research and the Privacy of Health Information. Beyond the HIPAA Privacy Rule: enhancing privacy, improving health through research [Internet]. Washington, DC: The National Academies Press; 2009 [cited 2015 Oct 9]. Available from: http://www.nap.edu/catalog/12458/beyond-the-hipaa-privacy-rule-enhancing-privacy-improving-health-through

23. Holman CD, Bass AJ, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust Health Rev. 2008 Nov;32(4):766-77.

24. Pencarrick Hertzman C, Meagher N, McGrail KM. Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest. J Am Med Inform Assoc. 2013 Jan;20(1):25–8.

25. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks CJ, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform. 2014 Aug;50:196–204.

26. Ventura M. Lei de acesso à informação, privacidade e a pesquisa em saúde. Cad Saude Publica. 2013 abr;29(4):636–8.

27. Silva M. Linkage de bases de dados identificadas em saúde: consentimento, privacidade e segurança da informação [tese]. Rio de Janeiro (RJ): Universidade Federal do Rio de Janeiro; 2012.

28. Almeida MF, Jorge MHPM. The use of the "linkage" technique of information systems in cohort studies on neonatal mortality. Rev Saude Publica. 1996 Apr;30(2):141–7.

29. Ministério da Saúde (BR). Secretaria de Vigilância em Saúde. Departamento de DST, Aids e Hepatites Virais. Boletim Epidemiológico HIV Aids. Brasília: Ministério da Saúde; 2014. Ano 3 n. 1