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

doi: 10.5123/S1679-49742015000400023

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

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.

* Claudia Medina Coeli receives a research fellowship from CNPq (304101/2011-7) and from Faperj (E26/102.771/2012). Rejane Sobrino Pinheiro receives a research fellowship from CNPq (309728/2012-6). Kenneth Rochel de Camargo Jr. receives a research fellowship from CNPq (300686/2013-7) and from Faperj (E-26/102.900/2012)

Correspondence:
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

Introduction

In recent years there has been an unprecedented expansion of the body of data that can be used in health research, surveillance and evaluation. This growth is the result of increased use of the Internet to access information, participation in social networks and use of Web applications, especially via mobile devices. Terms such as big data, data mining, text mining and web mining have become part of the population health research vocabulary[1]. The greater volume and variety of data, together with the application of techniques for extracting knowledge, can provide contributions to health research and evaluation. Nevertheless, there are some problems that may limit the potential gains from using this rich body of data. The selected profile of users of services available on the Internet, managers of these services applying filters which are not publicized and vulnerability in identifying correlations between events explained by chance are issues that must be considered[2,3]. Initiatives such as, for instance, the Big Data to Knowledge (BD2K) program implemented by the National Institutes of Health (NIH-USA), seek to explore the potential of big data for population health research[4].

The big health data era is still at the initial stages[5]. Notwithstanding, the availability of epidemiological, administrative and clinical databases in digital format, as well as the availability of record linkage techniques which enables them to be linked together has, in recent decades, consolidated the use of secondary data in health research and evaluation. In this paper we present a discussion on record linkage techniques, looking at methodological and ethical aspects, as well as their application in building population records. Finally we reflect on the challenges for research based on the use of RL techniques in Brazil.

Record linkage

Record linkage (RL) is the process of combining records for the same individual contained in two distinct databases, or identifying records that refer to the same individual in the same database[6].

An advantage of combining databases is that it enables aetiological hypotheses to be better explored using secondary data, given that it makes available a larger set of variables which in turn facilitate adjustment for confounding factors. Another advantage of applying RL is that it enables longitudinal, retrospective or prospective studies to be conducted[7]. Furthermore, the occurrence of health events during the course of an individual's life can be accompanied, from birth to death[8]. Finally, RL can also be used to improve the quality of health databases by enabling the identification and elimination of duplicated records, the filling in of missing data, the correction of fields containing invalid records, as well as correcting underreporting[9].

> *The availability of epidemiological, administrative and clinical databases in digital format, as well as the availability of record linkage techniques which enables them to be linked together has, in recent decades, consolidated the use of secondary data in health research and evaluation.*

RL Techniques

When the databases to be linked have a unique identifier, RL occurs in a relatively simple manner, and the majority of database management programs and statistical packages have commands for linkage based on a key identifier field. On the other hand, the majority of databases of interest for health studies do not have a common unique identifier. This means that it is necessary to use multiple fields holding personal data. In this case, RL involves several stages and different strategies can be used in each of them. A comprehensive and up to date discussion about the different strategies available can be found in the book written by Peter Christen[6]. What follows is a summary of the main stages of RL.

The first stage is database pre-processing, which covers data cleansing, standardization of contents and formats, breaking down name and address fields into their components (parsing), and the creation of phonetic codes. Pre-processing is performed only once in each database, and the resulting cleansed and standardized databases can be used in future data linking projects.

The next stage is indexing or blocking, which aims to reduce the number of record links that will be sent for comparison and classification. The traditional process consists of partitioning the two files based on the values of one or more fields used as the indexing key (for example, sex + phonetic code of first name + phonetic code of last name). Record comparison is thus restricted to records

in agreement with the value of this key. Indexing is a trade-off between comparing a large number of records and vulnerability to losing true links. With the aim of minimizing such losses, database linkage projects take place in multiple steps. In each of these steps an indexing key formed by a different combination of fields is used.

The links formed in the previous stage are processed in the comparison and classification stages. Exact or approximate functions are used for comparison and indicate similarity between two attributes (partial agreement). Comparison functions are specific for each type of data (for example, chains of characters, numbers, dates). For each record link the result relating to the similarity of each attribute is stored in a numerical vector. Finally, the similarity vector is used in the automatic classification process, which considers that the more two records are similar, the more likely it is that they refer to the same individual. The final result is the link classified as a true pair, a false pair or a doubtful pair. Doubtful pairs are sent for manual revision.

The most frequently used linkage techniques are the deterministic and probabilistic techniques based on the model created by Fellig & Sunter[10]. Deterministic techniques use exact comparison functions and a rule-based classification approach. The rules are developed based on specialized knowledge and specific computer routines need to be developed for each project. The probabilistic technique uses approximate comparison functions. Different weights are attributed to each field based on their discrimination power and vulnerability to error, whereby a score is calculated that indicates how true it is that two records belong to the same individual. It is an overall solution that can be applied to different linkage projects. A series of open source and commercial softwares implement this technique[6], so that special programming skills are not needed. The OpenReclink software (http://reclink.sourceforge.net/)[11], developed by us, implements Fellig & Sunter's model and is an evolution of Reclink software[12]. The open version, apart from being open source, includes novelties such as being multiplatform, multilanguage and having a revised version of the deduplication routine. The Linux version is still in current use in our laboratory at the Institute of Collective Health Studies (IESC-UFRJ) (LinkDataPop) in large-scale projects and offers good performance and stability. An alternative to the models described above is the use of machine learning-based routines. This model is implemented using ChoiceMaker software (http://oscmt.sourceforge.net/), which is adopted by The Centre for Health Record Linkage (CheReL) (NSW, Australia)[13]. The classification model is based on the use of rules which are weighted by applying machine learning techniques. The weight of each rule is used to calculate a score that indicates how true it is that two records belong to the same individual.

Figure 1 shows the initial screen of OpenReclink.

Regardless of the model adopted, the subset of links classified as doubtful in the automatic classification stage will need to be submitted to manual revision so that final classification can be performed. This is the stage that requires the largest amount of human resources and time to be completed. For example, in a study we conducted with the aim of identifying the number of previous live-born children of women who gave birth to live-born babies in the state of Rio de Janeiro in 2007 and 2008, a database holding around 2,400,000 records was processed and the revision process took some 800 hours. In order to optimize the manual revision process, based on the experience of revision experts, a routine for the automatic classification of links was developed based on twenty distinct criteria, generating a final score for each link. The algorithm was tested in an application involving linkage between the AIDS Notification System (SINAN-AIDS) and the Mortality System, and achieved good performance[14].

The last stage in a database linkage project should involve evaluation of the quality of the process in terms of the proportion of unmatched pairs, despite belonging to the same individual, and the proportion of false pairs, i.e., pairs that refer to different individuals. This stage is still not particularly valued in Brazilian studies using record linkage techniques[15]. Greater attention needs to be paid to the evaluation of linkage processes performed in Brazil, given that address information recorded on our databases does not always allow linkage to be used in automatic processing, and some surnames (e.g. Silva) and first names (e.g. Maria) are very frequent in Brazil[16], thus increasing vulnerability to the formation of false-positive pairs in the automatic classification stage.

Figure 1 – Home screen of the OpenReclink program with process menu at the top

Ethics and privacy

Ever since the publication of the seminal papers on RL by Newcombe et al.[17] and Fellig & Sunter[18] at the end of the 1950s and 1960s, respectively, new information technologies have been introduced and this has enabled a huge increase in database storage and processing capacity. The above mentioned book by Peter Christen[6] discusses several solutions capable of optimizing RL in terms of processing time and accuracy. As such, the main barrier to applying RL techniques in health research and evaluation is not technical but rather ethical, given that the majority of RL processes require access to personal data.

Several countries worldwide have legislation aimed at individual protection with regard to the use of personal data. This legislation determines that access to data can only occur when the individual in question gives their authorization. However, in recognition of the need to find a balance between ensuring individual rights and potential gains for society as a whole, in general this legislation treats the use of personal data in research as an exception and allows access without consent, provided the research project meets a series of requirements. One such requirement is that research be conducted in accordance with ethical guidelines and be approved by a research ethics committee. The ethical guidelines of the Council for International Organizations of Medical Sciences (CIOMS)[19] regarding epidemiological studies provide for the waiver of the requirement to obtain free and informed consent when: the research question is relevant for society; research subjects would be exposed to no more than minimal risk; researchers ensure that individual and collective rights will not be violated; norms are adopted to ensure privacy, confidentiality and anonymity; and when requiring consent would make the conduct of the study impracticable because of the difficulty in applying it.

In a systematic review into consent related to RL, we found a high proportion of requirement for consent in the majority of the studies reviewed[20]. However, in almost all cases consent was required for studies

involving primary data collection. Experience with requesting consent from the population in general such as, for example, organ donation programmes, suggests that implementing it is more problematic and that it may even be influenced by the type of question asked to request it (explicit consent – opt in, versus implicit consent – opt out)[21]. A report prepared by the Committee on Health Research and the Privacy of Health Information/ Institute of Medicine[22] recommends that the request for informed consent for the use of information for the purposes of health research and evaluation should be replaced by greater emphasis on measures to ensure transparency in the use, security and confidentiality of information. Countries such as Australia[23], Canada[24] and Wales[25] have consolidated experience in implanting data linkage units that provide a safe environment for storing and processing identified data, whilst also adopting data governance models that ensure balance between preserving the right to privacy and the potential gains of the use of data in population evaluation and research.

In 2011 the Access to Information Law (Law No. 12527) was enacted in Brazil[28], the purpose of which is to regulate the right to access public information. With regard to personal information, the law stipulates that access may only take place when the person consents. The law does however allow the waiver of the need to obtain consent in some situations, including scientific research. In July 2015, the period of public consultation on the preliminary draft of a law to protect personal data ended (http://pensando.mj.gov.br/dadospessoais/). The preliminary draft also states that scientific research should be treated as an exception with regard to the need for consent, although it does not go into further details, and this will probably be done by means of complementary regulations. In addition, Chapter IV, Item 8 of National Health Council Resolution 466/12, which regulates research involving human beings (http://conselho.saude.gov.br/resolucoes/2012/Reso466.pdf), admits the waiver of consent: "In cases in which it is impracticable to obtain Free and Informed Consent or when obtaining it would mean substantial risks to the privacy and confidentiality of the participant's data or to the bonds of trust between researchers and research subjects, a duly justified request for the waiver of Free and Informed Consent must be submitted by the principal investigator to the Research Ethics Committee/National Research Ethics Commission System for appreciation, without prejudice to the subsequent process of clarification". In a manner similar to that of other countries, while on the one hand Brazilian legislation seeks to ensure the right of the individual to privacy, on the other hand it allows a waiver of consent in research situations, thus seeking to find balance between the need to protect individual rights and the potential gains for society as a whole.

Data Linkage Units

One of the most interesting applications of RL is its use in building health records. These records are formed through the routine integration of health databases (vital statistics, administrative data, clinical data, for instance) with databases of other sectors (education, for example). Once they have been built, the records enable the generation of specific databases, in an efficient and privacy preserving manner, to answer questions of interest to health research and evaluation. In addition, they allow collected primary data to be integrated with record databases, assisting longitudinal follow-up of research participants, among other applications. Records are implanted in units equipped with infrastructure that ensures the processing and safekeeping of large volumes of identified data. In addition, the governance of data by all these units is done in accordance with publicized protocols.

Australia, Canada, Wales and Scotland are countries that have more consolidated experiences with this model of data linkage units. Australia has been implementing a national network of data linkage units since 2007 (The Population Health Research Network – PHRN) (http://www.phrn.org.au/about-us/overview). More recently a network formed by four data linkage units has been implanted in the United Kingdom, known as the "The Farr Institute of Health Informatics Research" (http://www.farrinstitute.org/). Although the model of these units varies in some aspects[26], generally speaking all of them share the mission of integrating databases that are diverse in their nature and making unidentified derived databases available for the purposes of evaluation and research. Moreover, their mission also includes developing techniques for secondary data processing and analysis, as well as training health technicians and researchers regarding RL issues and secondary data analysis.

The International Health Data Linkage Network - IHDLN (http://www.ipdln.org/) was created in 2008. The network aims to encourage cooperation between researchers and data linkage units in areas such as information technology, RL and secondary data analysis. In order to make clearer the kind of research done by network members, in 2014 it changed its name to "The International Population Data Linkage Network (IPDLN)". Twenty-three data linkage units are currently members of the network and are located in Australia (9), Canada (7), United Kingdom (5), Germany (1) and New Zealand (1). When the researcher's country of origin is taken into consideration, however, the network's geographical representativeness is greater (34 countries). At the time this paper was written, Brazil was the only South American country with network members.

Brazilian RL experience

Papers on health RL in Brazil began to be published in the mid 1990s[27]. Since then production has been increasing both in terms of the number and the variety of authors and the institutions of which they are members. A search conducted on the PubMed database on 11/10/2015, using the keyword "record linkage AND (*Brasil* OR Brazil) " found 68 papers. It is important to consider, however, that this number does not reflect total production of papers, given that many of them have been published in periodicals that are not indexed on the Pubmed database. An example is the thematic edition of the *Cadernos de Saúde Coletiva* periodical which was dedicated to RL (http://www.cadernos.iesc.ufrj.br/cadernos/index.php/features-sp-417739839/2006/no2-abr—jun).

Alongside the interest in RL applications in research, health service managers have also been interested in incorporating RL to reduce the underreporting of events and to improve the quality of health databases. The partnership between researchers and service managers at federal, state and municipal level has occurred through support for projects aimed at developing RL solutions to assist health surveillance and monitoring, as well as training health technicians in RL matters. A positive example of this partnership was the adoption, effective from 2004, of RL techniques to correct AIDS case underreporting on the SINAN notification system[28].

In view of the countless possibilities for RL use and its potential to contribute to the enhancement of health information quality, along with recent progress with the legal and regulatory framework (the Access to Information Law and National Health Council Resolution 466/2012), the way in which RL research is being conducted in Brazil has to change. Brazil needs to adopt a national model for linking population databases. International experience, based on data linkage units, can serve as a reference for building the Brazilian model. Nevertheless, it is fundamental that issues such as legislation specificities, experience in producing and disseminating secondary data, as well as the large volume and national scope of Brazilian databases be taken into consideration. As such, it is important to create mechanisms that allow researchers, service managers, data custodians and civil society representatives to work together so that the international model can be adapted to Brazilian specificities. One initiative in this direction was the "Seminar on Health Database Linkage" held in Rio de Janeiro in October 2014. The seminar was organized by the Ministry of Health's Information Technology Department (*DATASUS*) and was attended by stakeholders involved in the process of health data generation, custody and analysis, as well as guest participants from countries with outstanding development in this area (Canada, Wales and Australia). It is important that this initiative generates a follow-up, in order to define the legal and operational reference framework for database linkage activities in Brazil, both for research and service management.

However, in recognition of the fact that defining an institutional model for such a complex issue will take time, it is fundamental that a set of transition rules is established as soon as possible regarding access to identified databases that in addition to being compatible with prevailing legislation also allow projects under way to continue operating until a new model becomes available, so as to ensure that the progress and gains achieved with RL-based research in Brazil so far will not be lost. Be that as it may, simply maintaining the status quo is undesirable, given that some practices in current use thus far may be in conflict with the new legislation on access to data.

References

1. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Comput Biol. 2012 Jul 26;8(7):e1002616.
2. Ruths D, Pfeffer J. Social media for large studies of behavior. Science. 2014 Nov 28;346(6213):1063–4.
3. Khoury MJ, Ioannidis JPA. Medicine. Big data meets public health. Science. 2014 Nov 28;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. Filho C, Porto AD. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol Serv Saude. 2015 jun;24(2):325–32.
6. Christen P. Data matching concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin; New York: Springer; 2012.
7. Vidal EIO, Coeli CM, Pinheiro RS, Camargo 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 J. 2006 Oct;17(10):1569–76.
8. Kramer MR, Dunlop AL, Hogue CJR. 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 de, Pinheiro RS, Vasconcelos AMN, Bartholomay P, Oliveira GP de, et al. 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; 2007.
11. Camargo Jr KR de, Coeli CM. Going open source: some lessons learned from the development of OpenRecLink. Cad Saude Publica. 2015 Feb;31(2):257–63.
12. Camargo KR Jr, Coeli CM. Reclink: an application for database linkage implementing the probabilistic record linkage method. Cad Saude Pública. 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 25;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, Coeli CM. A qualidade do linkage de dados precisa de mais atenção. Cad Saude Publica. 2015 Jul;31(7):1349–50.
16. Coeli CM, Jr C, De KR. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. Rev Bras Epidemiol. 2002 ago;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, World Health Organization. International ethical guidelines on epidemiological studies. Geneva: World Health Organization; 2009.
20. Silva MEM, Coeli CM, Ventura M, Palacios M, Magnanini MMF, Camargo TMCR, 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 J. Nass, Laura A. Levit, and Lawrence O. Gostin, editors; Committee on Health Research and the Privacy of Health Information. Beyond the HIPAA. Privacy rule: enhancing privacy, improving health through research [Internet]. Washington: National Academy of Sciences; 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 CDJ, Bass JA, 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;32(4):766.
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 MH. 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. 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