Faculty of Health Sciences

Office of Research and Development

Enabling health research using administrative data: methodological improvements

Sean Randall

This thesis is presented for the degree of

Doctor of Philosophy (Supplication)

of

Curtin University

February 2017

Declaration

To the best of my knowledge this thesis contains no material previously published by any other person except where due acknowledgement has been made.

This thesis currently contains no material which has been accepted for the award of any other degree or diploma in any university.

Sean Randall

21 February 2017

Acknowledgement

I would like to thank my supervisors, Associate Professor Anna Ferrante and Professor James Semmens, for the time and effort they put into reviewing this manuscript. I would especially like to thank Associate Professor James Boyd and Associate Professor Anna Ferrante for their encouragement and mentorship over all these years; it has been invaluable.

Abstract

The amount of data collected about individuals is increasing dramatically [1]. This data has enormous potential for secondary use, particularly for research. Record linkage is the process of joining together datasets to determine which records within and between datasets convey information about the same individual. Record linkage is a key epidemiological tool, allowing researchers to answer detailed questions about the health of entire populations at low cost. Linked health data has been used to answer questions on the nature of diseases (such as estimates of disease prevalence, incidence and survival), the impact of introduced treatments and the impact of health policy. Research using linked data has given a greater understanding of the nature of disease and led to changes in health service delivery and policy [2]. Australia has long been a leader in this area, with dedicated linkage infrastructure first established in 1995. Further investment in Australia's record linkage capability began in 2009 through the National Collaborative Research Infrastructure Strategy initiative with the allocation of over $100 million from federal and state governments to develop record linkage infrastructure across the country [3].

There are two key methodological challenges for record linkage as practiced in Australia; quality, and privacy.

Epidemiologists require high record linkage quality to ensure accurate results. However, current methods to ensure quality are typically manual, and thus costly, time consuming and difficult to scale. As dataset sizes grow, these methods quickly become untenable – new methods are required to ensure high linkage quality at low cost. The first aim of this thesis is to evaluate current practice and develop new methods to improve linkage quality.

Health data is sensitive information, and care must be taken to ensure this data is kept confidential and the privacy of individuals is protected. Despite the level of privacy offered by current record linkage arrangements, many data custodians are either legally constrained or do not feel comfortable with the risks associated with allowing their data to be used for record linkage. Methodological improvements have been suggested in the record linkage literature which would allow linkage to occur on encrypted data, known as privacy preserving record linkage. Methods such as these may prove more palatable to constrained or reluctant custodians. The second aim of this thesis is to develop and evaluate practical methods for privacy preserving record linkage.

This PhD presents methodological advancements in record linkage to improve both linkage quality and privacy protection. Along with these advances, an overview of recent record linkage developments in Australia is presented, outlining the establishment of national data linkage infrastructure along with examples of research using linked data.

Several areas of record linkage processing are investigated to determine best practices for ensuring high record linkage quality. Novel methods for identifying errors in linkage are presented and evaluated. A method for privacy preserving record linkage which requires only encrypted personally identifying information is also introduced and evaluated. This method is shown to produce linkage quality (accuracy) that is comparable to current methods using full personal identifiers. This method is further compared to techniques currently in use in Australia to preserve privacy. Finally, a highly secure privacy preserving protocol is presented utilising newly developed cryptographic algorithms.

Centre for Population Health Research Director's recommendation

Thesis content for assessment

The National Collaborative Research Strategy (NCRIS) was initiated in Australia in 2009 and supported the development of national data linkage infrastructure as a priority theme within the Population Health Research Network (PHRN). Within this initiative, the Centre for Data Linkage (CDL) was established within the Centre for Population Health Research (CPHR) at Curtin University to provide cross-jurisdictional national data linkage capability between the States and Territories of Australia. To meet the needs of this strategic initiative, a high quality team was assembled to establish this national data linkage infrastructure and to support the development of international data linkage collaborations in the United Kingdom, northern Europe and Canada.

Mr Sean Randall was employed as part of this expert team. Sean has worked closely with Associate Professors James Boyd (CDL Director) and Anna Ferrante (CDL Deputy Director) and has provided a significant role in the general development of the national data linkage framework, as evidenced by ten supporting papers. In addition, Sean has provided an expert lead role in developing methodologies in 'privacy-based linkage' and 'linkage quality' within this framework. His significant expertise and contribution in these areas is demonstrated by six scientific publications where he is the lead author and are discussed in Chapters 4 and 5. It should also be noted that Sean provided the lead role in the implementation and refinement of Privacy Preserving Record Linkage (PPRL) methods which has gained significant international interest and is currently the basis of several international collaborations involving data linkage centres in Australia (Western Australia, New South Wales), Germany, the United Kingdom (through the Farr Institute in Bristol, Wales, Scotland and northern Ireland), Canada (Institute for Clinical Evaluative Sciences, Toronto; and University of British Colombia). These collaborations are also being supported by a DAAD grant (German-Australian exchange scheme), the NCRIS PHRN (PPRL was selected as a 2016/17 Strategic Priority project) and a current submission to the Canadian Institutes of Health Research.

It should also be noted that it is standard practice in health research to have multiple authors included on a paper given the collaborative nature that supports these scientific endeavors.

Signed: Date: 21 February 2017

(Professor James Semmens; Director, Centre for Population Health Research, Curtin University)

Overview

Record linkage is the practice of identifying which person-based records, both from within a single dataset, or across multiple datasets, belong to the same individual. In circumstances where a unique identifier exists (for instance Denmark has a national personal identifier required by all citizens to access basic services [4]) this process is relatively straightforward, with the records simply joined together through this unique identifier. When unique identifiers are not available, personally identifying information such as name, date of birth and address are typically used, which may be partially missing, in error, or change over time.

Record linkage is an important technique for observational research. Widely used in the health sector, it enables hospital, emergency and primary care collections, as well as birth, death and disease registries to all be joined together to create an overall picture of the health of an individual over time. This provides researchers with a cost effective, longitudinal resource for the entire population.

Australia has long been at the forefront of developments in linkage infrastructure. Western Australia began operating a dedicated linkage unit in 1995, linking health data within the state into a central repository to be used by future research projects (previously data had been linked for specific research projects only) [5]. This repository model of on-going record linkage dramatically increased access to linked data, although it also significantly increased technical complexity; the challenges of a repository model are discussed in *Publication 1* (see publication list p15-17) of this thesis.

Since 2009, there has been significant government investment in data linkage infrastructure in Australia [3], with the establishment of the Population Health Research Network (PHRN) under the National Collaborative Research Infrastructure Strategy. A key component of the PHRN was the ability to link together data from more than one state or territory. This is necessary due to the federated nature of healthcare in Australia, where some services are provided by states and territories and others by the federal government. The establishment in Australia of infrastructure for cross-jurisdictional research is described further in *Publication 2* of this thesis.

The PHRN initiated a number of 'proof of concept' projects to demonstrate the feasibility of linking data from across the country together to answer nationally important research questions. The first of these projects linked over 44 million records from across four Australian states; this project is described further in *Publication 3* of this thesis.

The development of national linkage infrastructure within Australia presents enormous opportunities for health research. The crucial advantage of linked data is the ability to answer questions about entire populations without the prohibitive cost this would typically entail. The use of linked data for health research is only limited by the nature of the collected information. Linked data has been used to answer questions on the nature of diseases (such as estimates of disease prevalence, incidence and survival), the impact of introduced treatments and the effectiveness of changes in health policy. This thesis presents several examples of research using linked data. Firstly, linked data can be used to gain a more accurate picture of health trends over time, as shown in *Publication 4,* where a more accurate understanding was gained of trends in the incidence of acute myocardial infarction.

Linked data can also be used to generate new hypotheses and knowledge about health conditions. For example, a research program into the long term effects of burn injury was initiated due to the documented persistence of inflammatory responses after burn injury. Using linked data, it was found that those with burn injury have higher long term mortality, and higher rates of cardiovascular and musculoskeletal disease (see *Publications 5-7).* Further experimental research has subsequently confirmed that pathophysiological changes are the likely cause [6]. The use of linked data here played a large part in the general of new knowledge about burn injury.

To enable this health research, robust linkage methods are required, which can ensure quality and reduce privacy risk. The core of this thesis consists of a series of publications addressing linkage quality and privacy, respectively, which are addressed in Chapters 4 and 5.

Ensuring linkage quality

Researchers require high linkage quality to ensure the accuracy of their research. Linkage quality here refers to a low number of false positives (records which are designated as belonging to the same person when, in truth, they do not) as well as a low number of false negatives (records which are designated to separate individuals when, in truth, they actually belong to the same person). There is little understanding of the direct effect that linkage errors have on research results, although linkage error is more likely to occur in vulnerable populations [7]. These issues are discussed further in *Publication 8*.

Efforts to optimise linkage quality play a large role in linkage operations. Current methods can involve manual review of record-pairs to ensure quality [8, 9], a costly and time consuming process. By determining best practices for record linkage operations with regard to linkage quality, and by developing new methods to optimise quality and reduce the manual processing burden, this thesis aims to allow higher linkage quality to be achieved quickly and at lower cost.

To evaluate linkage quality in practice, methods are required to calculate it; one solution to this problem is the use of a statistical sampling methodology; a method which provides accurate results is presented in *Publication 9*.

The record linkage process, of which the comparison of individual records to determine whether they belong to the same individual is but one step, is outlined below (see Figure 1).

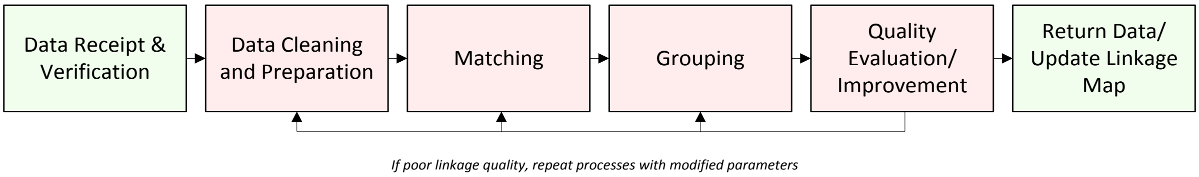

Data Receipt & Verification
Data Cleaning and Preparation
Matching
Grouping
Quality Evaluation/ Improvement
Return Data/ Update Linkage Map
If poor linkage quality, repeat processes with modified parameters

Figure 1: The record linkage process

Upon receipt of data and the verification of its contents, a data cleaning and preparation stage typically occurs in order to standardise the file as required and perform any necessary data cleaning (explained in further detail below). Matching then occurs, whereby individual records are compared against each other to determine whether the pair of records belongs to the same individual. The most common method used for matching is known as probabilistic record linkage, so called because it uses conditional probabilities to determine the likelihood of two records belonging to the same person. The subsequent grouping process amalgamates the record-pairs generated through matching into a *linkage map*, listing all the records belonging to each individual [10]. All of these processes involve making numerous decision about the most appropriate parameter settings to achieve high linkage quality.

A final step is one of performing an evaluation of the created linkage map, by automated or manual methods, to determine its accuracy or quality. This evaluation processes may be used to directly improve the linkage map quality (as in manual review, where groups of records are manually edited and the results saved into the linkage map), or they may be used to evaluate whether the entire record linkage process should be re-started, using alternate parameter settings, for instance. Once an acceptable standard of linkage quality is reached, the record linkage process is complete.

As shown above, linkage processing is made up of a number of discrete parts, with each influencing overall linkage quality. The publications in this section have focused on particular record linkage processes, aiming to evaluate current practice and develop new methods to improve overall quality.

Data cleaning involves the transformation of the information received for linkage into a standardised format that is most appropriate for matching. The purpose of this process is to improve overall linkage quality. Despite widespread use, the effect of data cleaning on linkage quality has not been previously evaluated, and it is not known which techniques yield the most improvement. *Publication 10* outlines methods for data cleaning and their prevalence, and evaluates the impact of these methods on linkage quality, using both synthetic and real administrative data. The results suggest that rather than lead to noticeable improvements in linkage quality, heavy data cleaning can reduce the overall quality of linkage.

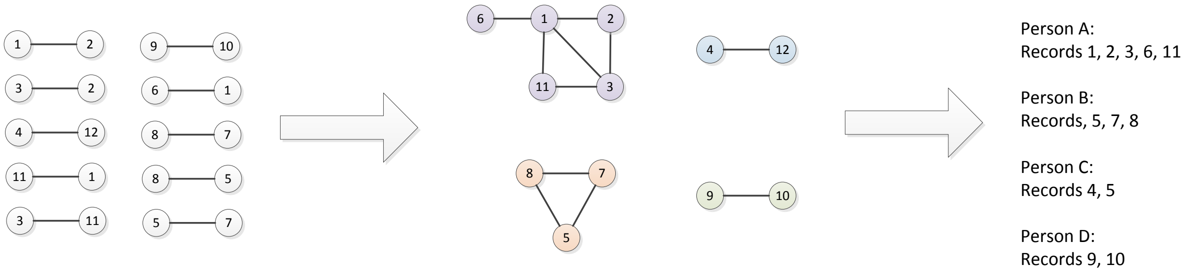
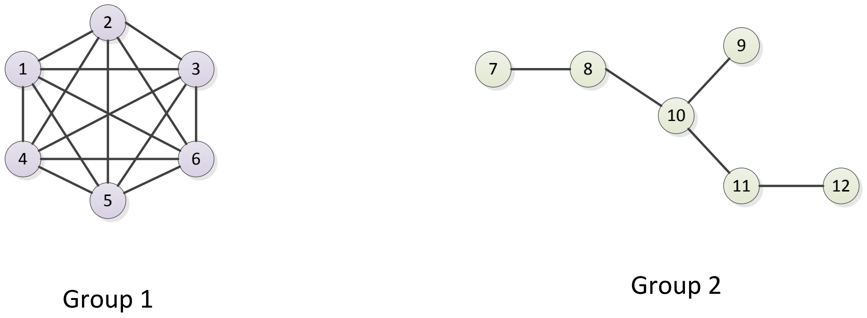


Figure 2: The merge grouping process

While the matching process results in a set of record-pairs thought to belong to the same individual, the grouping process converts these record-pairs into a list outlining which records belong to which individual - a linkage map. The standard grouping strategy is to amalgamate all pairs above an excepted threshold, as shown in Figure 2. Alternate methods can be used in specific linkage scenarios, where additional information is known about the composition of each dataset. Organisations which conduct record linkage on a regular basis typically use a repository model, whereby one single linkage map is created from all provided datasets, which is refreshed as new and updated datasets are added. In such cases, the organisation may not wish to have new data merge together groups within the repository, as there is high confidence that existing groups in the repository represent separate people. *Publication 11* describes and evaluates both existing and novel methods for grouping with a repository of previously linked data. Results suggest that alternate approaches achieve substantially better quality.

The final publication in the section on linkage quality focuses on methods to improve linkage quality after linkage is complete. Typically human manual review is used; however, this is slow and infeasible for large datasets.

The amalgamation of record-pairs into a linkage map through the grouping process provides information unavailable during linkage that has potential to be useful for improving linkage quality. Groups of records which are more sparsely held together may be more likely to contain false positive links, as compared with groups that are fully saturated with pairs (an example is shown in Figure 3). *Publication 12* uses measures from the mathematical field of graph theory to identify groups of records likely to contain errors. These measures accurately identified groups containing errors with superior precision to the typically used threshold setting methods.



Group 1

Group 2

Figure 3: Is Group 2 more likely to be in error than Group 1?

The publications in this section aim to improve the quality of record linkage without resort to largely manual, and thus expensive, techniques. By increasing overall linkage quality, the confidence in researcher results can in turn be increased.

Ensuring privacy

Administrative health data is highly sensitive. It contains personal information about individuals which could cause harm if it became publically known. Health data can be considered the most intimate and personal of any information routinely collected about an individual [11].

Record linkage units operate under strict privacy arrangements and take great care to manage the confidentiality of their data. It is vital that adherence to rigorous privacy protection mechanisms occur to ensure public confidence. Privacy issues continue to have a large impact on the extent and quality of linked health research, with important data collections underutilised due to the privacy concerns of their custodians [12]. Efforts to guarantee privacy include governance and information technology controls, the utilisation of specific data flows, and the development of privacy preserving record linkage methods. These are discussed further in *Publication 13.*

In privacy preserving record linkage, personal identifiers are encoded or encrypted before being provided to any third party. The specific encoding or encryption used allows record linkage to still occur. This method of linkage has much less privacy risk, as no personal identifiers are released to third parties.

Research into this method of linkage is ongoing, with an array of new protocols in the literature [13]. The most promising of these, preserving record linkage using

Bloom filters [14], allows approximate matching to occur, a technique important for ensuring high linkage quality. *Publication 14* proposes extensions to this method to allow full probabilistic record linkage, and evaluates the protocol on large-scale administrative data. The presented protocol achieves results equal to that achieved using unencrypted identifiers in a similar time frame. Given these results, this privacy preserving linkage approach appears a genuine alternative to standard unencrypted record linkage.

Publication 15 compares this privacy preserving approach with another which have received some use in Australia; the Statistical Linkage Key-581 (SLK) [15]. This 'key' is created by amalgamating specific letters from a person's first and last names along with their sex and date of birth into a single field. This field is taken to be unique; all records with the same SLK are identified as belonging to the same individual. While the SLK is made up of identifiable information, an identity is not immediately discernible from the key. In *Publication 15,* the Bloom filter method is shown to achieve higher linkage quality, and provide greater privacy than the SLK method.

Recent analysis of the Bloom filter method has shown that in some circumstances the method may be vulnerable to frequency attacks [16]. Frequency attacks use the fact that certain identifiers occur more frequently than others (for instance, the first name 'John') to learn information about the encrypted data. While the Bloom filter method still provides superior privacy as compared to other practical alternatives, a method which improves upon its security would be favourable. *Publication 16* provides a new protocol which build upon the previous Bloom filter protocol, but provides greater security. It is impervious to frequency based attacks and achieves equal linkage quality to the previous Bloom filter method. Its shortcoming, however, is that it is slower.

The Bloom filter method presented here offers an important opportunity to advance the privacy of record linkage. In doing so, the risks of data release are lowered; it is hoped that this advance can improve overall access to linked data, and thereby allow further important research questions to be answered.

List of peer reviewed publications included as part of this thesis:

Chapter 1: The history and development of record linkage

Supporting Publications

1. Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., Brown, A. P., & Semmens, J. B. (2014). Technical challenges of providing record linkage services for research. *BMC medical informatics and decision making, 14*(1), 1.

2. Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., **Randall, S. M.**, & Semmens, J. B. (2012). Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC health services research*, 12(1), 1.

3. Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., McInneny, K., Brown, A. P., Spilsbury, K., Gillies, M, & Semmens, J. B. (2015). Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia. *BMC health services research*, *15*(1), 1.

Chapter 2: Uses of record linkage

Supporting Publications

4. **Randall, S. M.**, Zilkens, R., Duke, J. M., & Boyd, J. H. (2016). Western Australia population trends in the incidence of acute myocardial infarction between 1993 and 2012. *International Journal of Cardiology, 222*, 678-682.

5. Duke, J. M., Rea, S., Boyd, J. H., **Randall, S. M.**, & Wood, F. M. (2015). Mortality after burn injury in children: a 33-year population-based study. *Pediatrics, 135(*4), e903-e910.

6. Duke, J. M., Boyd, J. H., Rea, S., **Randall, S. M.**, & Wood, F. M. (2015). Long-term mortality among older adults with burn injury: a population-based study in Australia. *Bulletin of the World Health Organization, 93*(6), 400-406.

7. **Randall, S. M.**, Fear, M. W., Wood, F. M., Rea, S., Boyd, J. H., & Duke, J. M. (2015). Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study. *BMJ open*, *5*(9), e009395.

Chapter 4: Methods for improving quality

Supporting Publications

8. Boyd, J. H., Ferrante, A. M., Irvine, K., Smith, M., Moore, E., Brown, A. P., **Randall, S. M.** (2016). Understanding the origins of record linkage errors and how they affect research outcomes. *Australian and New Zealand Journal of Public Health, In Press*

9. Boyd, J. H., Guiver, T., **Randall, S. M.**, Ferrante, A. M., Semmens, J. B., Anderson, P., & Dickinson, T. (2016). A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. *Methods of information in medicine*, *55*(3), 276-283.

Key publications

10. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making, 13*(1), 64.

11. **Randall, S. M.**, Boyd, J. H., Ferrante, A. M., Bauer, J. K., & Semmens, J. B. (2014). Use of graph theory measures to identify errors in record linkage. *Computer methods and programs in biomedicine, 115*(2), 55-63.

12. **Randall, S. M.**, Boyd, J. H., Ferrante, A. M., Brown, A. P., Semmens, J. B. (2015). Grouping methods for ongoing record linkage. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

Chapter 5: Methods for improving privacy

Supporting Publications

13. Boyd, J. H., **Randall, S. M.**, & Ferrante, A. M. (2015). Application of Privacy-Preserving Techniques in Operational Record Linkage Centres. In *Medical Data Privacy Handbook* (pp. 267-287). Springer International Publishing.

Key publications

14. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics, 50*, 205-212

15. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., Brown, A. P., Semmens, J. B. (2016). Limited privacy protection and poor sensitivity: is it time to move on from the Statistical Linkage Key-581? *Health information management journal 45*(2), 71-79.

16. **Randall, S. M.**, Brown, A. P., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2015). Privacy preserving record linkage using homomorphic encryption. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

Glossary

Agreement weight: The score given in probabilistic linkage when two identifiers (such as first name) match.

Best link grouping: A method of grouping used when one dataset should not bring two groups of records in a second dataset (or repository) together. It instead chooses the 'best' link.

Blocking: Any technique which seeks to reduce the number of comparisons required to be performed in linkage. Typically a field (i.e. surname) or a set of fields is chosen; only records with the same values of these fields are compared further.

Blocking variables: A field (i.e. surname) that is used for blocking.

Bloom filter: A data structure which can be used for privacy preserving record linkage. It allows for approximate matching.

Clerical review: see manual review

Content data: Data required by researchers, but not typically used (or provided) for linkage.

Data cleaning: The transformation of fields into a format best suited for linkage.

Data custodian: An 'owner' of the data. As owner, the custodian will necessarily have access to both the personal identifiers and the content data.

Data linkage: see record linkage

Deduplication: A linkage which aims to identify records within a single dataset which belong to the same individual (as opposed to linking together two different datasets).

De-identified data: Data that has had names and other personal identifying information removed such that it is no longer 'reasonably' identifiable.

Deterministic linkage: A method of linkage which utilises hard coded rules about when two records belonging to an individual.

Disagreement weight: The score given in probabilistic linkage when two identifiers (such as first name) do not match.

Entity resolution: Entity resolution seeks to determine entities (i.e. the same 'things') between and within datasets. In entity resolution, a single record may contain a number of different entities. Entity resolution can be considered a broader field within which record linkage is contained.

Epidemiology: The study of the distribution, causes and effects of health and disease.

F-measure: The harmonic mean between precision and recall, F-measure is used in this thesis as the main measure of linkage quality.

Field: A single personal identifier i.e. surname, postcode, sex.

Frequency attacks: A technique which utilises the frequency of letters, combinations of letters, and words to break or partially break an encryption scheme.

Graph theory: A branch of mathematics which study pairwise relations.

Grouping: This process whereby a list of pairs of records which belong to the same individual is converting into a linkage map.

Hash: An irreversible data transformation with a fixed size output. The word can refer to both the output of the algorithm (a hash), as well as the process of generating the output (to hash).

Homomorphic encryption: A method of encryption which allows simple mathematical operations to be applied to the encrypted values which, when decrypted, match the results of the same mathematical operations applied to the original unencrypted values.

Linkage map: A list of each record along with its corresponding person identifier. The development of a linkage map is the immediate purpose of record linkage.

Linkage quality: A measure of how many errors are found within a linkage. High linkage quality implies there are few errors. There are two types of errors possible for a linkage; false positives, where two records are designated to belong to the same person but do not, and false negatives, where two records are designated to belong to separate people, but do not.

Linkage strategy: The set of parameters used for a particular linkage.

Linkage unit: Organisation which conducts record linkage; typically independent from the data custodians and the researchers.

Machine learning: A field of computer science that aims to give computers the ability to learn how to complete a task without being explicitly programmed to do so.

Manual review: The process of determining whether records belong to the same individual by manually inspecting them, and making a decision using human intuition.

Merge grouping: The most common method for grouping. All record pairs above a specified threshold are amalgamated, with all connected records classified as belonging to the same individual.

M-probability: The probability that two records have the same value for a particular field, when the two records belong to the same person - used to determine agreement and disagreement weights in probabilistic linkage.

One to one linkage: A form of linkage between two datasets where each dataset has at most one record per person.

Parameter settings: Each method for linkage requires a number of parameters to be set, which will determine the results of the linkage.

Personal identifiers: The attributes of an individual commonly recorded in administrative data which can be used to identify them. The most common would be full name, sex, date of birth and address information.

Phonetic encoding: A method of encoding individual fields which seeks to ensure words which sound the same but have different spellings have the same representation.

Precision: A measure of quality, defined as the number of true positives divided by the sum of true positives and false positives. Precision provides a measures of the proportion of false positives.

Privacy preserving record linkage: Record linkage carried out on encrypted or encoded personal identifiers; in this process, the linkage unit has no access to full personal identifiers, only some form of encoded information.

Probabilistic linkage: A method of linkage which uses conditional probabilities to determine the likelihood of two records belong to the same individual based on the record attributes. Based around a formal statistical model, this method of linkage is the most common.

Recall: A measure of quality, defined as the number of true positives divided by the number of true matches (i.e. the sum of true positives and false negatives). Recall provides a measure of the proportion of false negatives.

Record linkage: Record linkage is the process of joining together administrative datasets to determine which records within and between datasets belong to the same individual.

Record-pairs: Pairs of records; the output of matching, record-pairs are two records thought to belong to the same person.

Repository (linkage map): An overarching linkage map which is updated over time. As different researchers often wish to use the same linked datasets for their own research, linkage units often adopt a repository model, where new data is linked into a maintained repository of links; Information is extracted from this overarching linkage map for individual researcher projects as required.

Separation principle: A method for improving privacy by restricting the type of data received by each organisation involved in linkage. The linkage unit receives the personally identifying information, but not the content data, while the researcher receives only the content data, and not the personally identifying information. Under the separation principle, only the data custodian has access to both personal identifying information, and clinical content data.

Statistical Linkage Key (SLK): A ‘key’ created by amalgamating specific letters from a person’s first and last names along with their sex and date of birth into a single field. This field is taken to be unique; all records with the same SLK are identified as belonging to the same individual.

String similarity/string comparison techniques: Methods to allow approximate comparison of two fields. Instead of two fields being designated a match if they completely agree, and not a match otherwise, some intermediate level is allowed. This is useful to handle spelling mistakes, for instance. A 'string' here is a computer science term for an alphabetic field.

Supervised (machine learning): Machine learning methods which require a set of training data. The machine learning data must first teach itself how to solve the problem on the training data, before it can be applied to new data.

Synthetic data: 'Made up' data, generated to allow testing and evaluation. Synthetic administrative data can be created with the answers regarding which record belongs to the same individual.

Threshold: A parameter in probabilistic record linkage, record-pairs are designated a match if their overall score is greater than a threshold.

Training data (machine learning): Data to be used in training a supervised machine learning method. This data must be as similar as possible to the data from the underlying problem the machine learning method is to tackle; however the training data must also contain the 'answers'.

Transitive closure: see merge grouping

Unsupervised (machine learning): Machine learning methods which do not require any training data.

U-probability: The probability that two records have the same value for a particular field, when the two records do not belong to the same person - used to determine agreement and disagreement weights in probabilistic linkage.

Table of contents

Aims and structure of this thesis

There are two overarching aims of this thesis.

Aim 1: Evaluate current practice and develop new methods to improve linkage quality

Aim 2: Develop and evaluate practical methods for privacy preserving record linkage.

Before tackling the above aims, the opening chapters of this thesis seek to establish the necessary background in this area. Chapter 1 outlines the history and development of record linkage processes, and provides supporting publications which address the development of record linkage in Australia and the shift towards ongoing linkage. Chapter 2 discusses the uses of record linkage, with supporting publications as examples. Chapter 3 presents a detailed review of the literature on record linkage methodology and helps to contextualise the material presented in Chapter 4 and 5. It aims to identify trends throughout the whole literature and provide suggestions for future work.

Chapters 4 and 5 contain the core of this thesis. Chapter 4 addresses linkage quality; its importance, how to measure it, and how to improve it. This chapter includes three key publications, along with additional supporting publications. In Chapter 5, privacy is discussed and methods to improve privacy protection in record linkage are presented in a set of three key publications, along with a supporting publication. The conclusion, in Chapter 6, discusses the research in context; its strengths, limitations and implications.

Chapter 1

The history and development of record linkage

Research Output

Supporting Publications

1. Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., Brown, A. P., & Semmens, J. B. (2014). **Technical challenges of providing record linkage services for research**. *BMC medical informatics and decision making, 14*(1), 1.

2. Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., **Randall, S. M.**, & Semmens, J. B. (2012). **Data linkage infrastructure for cross-jurisdictional health-related research in Australia.** *BMC health services research*, 12(1), 1.

3. Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., McInneny, K., Brown, A. P., Spilsbury, K., Gillies, M, & Semmens, J. B. (2015). **Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia**. *BMC health services research*, *15*(1), 1.

1.1. A history of record linkage

The use of administrative records to investigate aspects of disease has a long history, beginning with John Graunt's work investigating causes of death from the City of London's weekly *Bills of Mortality* [17]. The appeal of merging existing administrative records of an individual together was recognised since the 19th century (see Gill [10]) , however it was the adoption of information technology in the mid-20th century, and the move away from paper records to computerised methods of information storage, that began to make record linkage feasible.

Even the earliest papers on this new method of computerised record linkage note the technical challenges associated with the unreliability of identifying information; misspellings, swapped fields, missing data and incorrect data [18]. While humans can quickly judge whether two records are likely to belong to the same individual, concisely describing rules for all circumstances in algorithmic form was considered potentially infeasible [18]. Computers could dramatically speed up tasks such as sorting, vital for record linkage, but struggled with tasks intuitively handled by humans. Partly for this reason, manual clerical intervention remained a continuing part of record linkage practice [8, 19], despite its widely recognised cost [20, 21]. While computers make tasks such as record linkage feasible, the number of computations can quickly overwhelm available hardware, resulting in impractically long run-times. Despite the dramatic improvement in computer power over the last 60 years, issues of linkage quality and timeliness remain at the forefront of modern record linkage research [22].

The techniques developed in these early years of record linkage have remained an important part of modern record linkage practice - the use of blocking techniques to reduce the comparison space, the use of phonetic encoding techniques to remove typographical errors caused by similar sounding words and the use of probabilities to determine the likelihood of a match. The methods developed in these early papers were shown to be practical and provide high quality results, and were specifically designed to be applicable to different datasets containing different identifying information [23]. The calculation of frequencies of particular fields, to create agreement and disagreement probabilities in order to determine the overall likelihood of two records agreeing, later mathematically formalised [24], has become known as probabilistic record linkage, and is still the key theoretical basis for most record linkage work.

Despite the development of a robust record linkage methodology, in practice this was not always used, with ad-hoc in-house techniques common throughout the 1970s and 1980's. The apparent simplicity of the record linkage problem (and thus the simplicity of developing ad-hoc methods), as well as the lack of general purpose software implementing the more theoretical probabilistic approach, may account for this development [10]. These ad-hoc techniques eventually became known as *deterministic* record linkage. Despite the lack of a formal framework, in some circumstances they have been shown to work as well as probabilistic methods [25].

Early record linkage developments were concentrated in a few key locales. Canada was one of the earliest adopters of record linkage methods, with British Columbia linking together birth, marriage and death certificates for studies into radiation and genetic heritability [18]. The Oxford Record Linkage Study soon followed, utilising hospital inpatient information (including midwives records and stillbirth information) along with birth and death data [26] for a wide range of studies [27-29]. The US Bureau of Census was also involved in record linkage around this time, linking census and administrative records together to determine coverage, and releasing a record linkage software package in 1972 (*UniMatch, [30]).* Record linkage of health datasets existed on a national scale in Scotland from the late 1970's [31]. A search of the literature reveals record linkage for health research occurring during the 1980's in Iowa[32], California [33], Hawaii [34], Sweden [35], Finland [36] and Western Australia [37].

Several of these localities developed more comprehensive systems that others. Major record linkage systems were developed in Canada [38], Western Australia [39], and Scotland [40]. Each of these contained a sizable number of administrative collections along with the large population base necessary for epidemiological analysis. The Scandinavian countries also became significant sites for linked research; however, these sites had little need to develop complex systems to deal with the uncertainty of personal identifiers, given their use of state-issued unique person numbers [4, 41].

Around this time the ad-hoc linkages carried out by researchers to answer specific questions evolved into organisational structures known as *linkage units.* These linkage units were created to carry out linkages of large collections and store the results to be used by researchers who requested this information. The development of linkage units increased accessibility of linked data, and resulted in the fostering of specialist expertise in record linkage (previously epidemiologists had conducted linkages themselves). The development of linkage units saw the development of continuous 'linkage maps', with information on the relationships between records maintained over time as each new year of administrative records, once collected, is linked into the system. This development added further technical complexity. These technical challenges are discussed in detail in *Publication 1: Technical challenges of providing record linkage services for research* [42].

The ethical and legal implications of combining information on individuals (generally without direct consent), essentially a form of state surveillance albeit for noble purpose, was recognised as a significant challenge to record linkage practice from the outset [43]. The jurisdictions where record linkage operations grew were those without fixed legislative barriers. Techniques to manage and reduce privacy risk were established and implemented [44]; the development of these methods is an ongoing concern [13].

The introduction of record linkage for health research dramatically increased the usefulness of administrative record collections and resulted in thousands of publications on every facet of human health. The impact of this research on health policy has been well-documented [2]. Record linkage is now recognised as an important tool in modern epidemiology, featuring in both textbooks [45-47] and as part of academic coursework [48-50].

1.2. Record linkage in Australia

As early as 1970 Australia was recognised as an area particularly suited to epidemiological research using administrative data, given the number of high quality collections [43]. However the federated nature of Australian data collections, along with legal and privacy issues, were also recognised as potential challenges.

Early record linkage studies in Australia were conducted to answer specific research questions; examples include the investigation of industrial exposure to asbestos and lung disease, which involved linking employment records with cancer registries, hospital morbidity systems and death registries [51]; the linkage of criminal justice datasets for longitudinal research into offenders [52] and the development of a linked database for analysing maternal and child health, using perinatal records, hospital admissions records and birth and death registry information among others [53]. In 1995, work began to link health data within Western Australia into a central repository to be used by future research projects. The Western Australia Data Linkage Branch (WA-DLB) [39] initially included records from birth, death, hospital, mental health, midwives and cancer registry collections over a 15 year period.

The development of this database, along with a client services component providing a mechanism for access to this resource, has been a huge boon to health researchers. Since 1995, the WA-DLB has provided data for more than 700 research projects, resulting in both an enormous number of publications, and changes to health policy and clinical practice [54]. The linkage system has also increased in size, containing over 88 million records with over 4 million individuals [55]. Over 400 data collections have been linked using the infrastructure [55].

Building on from the success of the WA-DLB, the Centre for Health Record Linkage (CHeReL) was established in 2006 to conduct record linkage of NSW datasets [8]. By the end of 2015, the CHeReL linkage system contained over 94 million records and 11 million individuals [56], and had completed over 120 projects [57]. As with WA-DLB, these are records from state-based health datasets; many areas of health governed by the Commonwealth (such as primary care, and subsidised medicines) having rarely been made available for use in linked research.

In 2009, Australian state and federal governments, along with universities and other research centres, invested significantly in record linkage infrastructure through the National Collaborative Research Infrastructure Strategy. This provided resources for the establishment of record linkage units in several jurisdictions (South Australia, Queensland, Victoria and Tasmania), as well as for a centre for national and cross-jurisdictional linkage. It also provided funding for the establishment of a remote access laboratory, providing researchers a secure, remote environment in which they can conduct their research [58]. These organisations, together with the previously established WA-DLB, CHeReL and several others, formed the Population Health Research Network (PHRN). The creation of a cross-jurisdictional linkage capability (the ability to link together data from more than one state or territory) was a key component of the PHRN.

The establishment of this infrastructure for cross-jurisdictional research, along with the processes and methodology involved, are described further in *Publication 2: Data linkage infrastructure for cross-jurisdictional health-related research in Australia [59].*

Cross-jurisdictional linkage has numerous benefits. It can provide a picture of the nation as a whole, rather than just individual states. It allows the combination of federal health datasets (such as primary care and prescribing information) with state based health information (such as hospital and emergency datasets). Health issues of common interest between states can be investigated (such as health service utilisation in towns which exist across state borders). The loss to follow up caused by interstate movement, or service provision across state borders is reduced. The larger population increases statistical power, which provides additional opportunities to investigate rare conditions or rare outcomes.

The PHRN initiated several 'proof of concept' collaboration project to establish the feasibility of conducting cross-jurisdictional linkage to answer research questions of national importance. The first of these proof of concept projects linked over 44 million records from across four Australian states to form a national linkage map, the first of its kind in Australia. This proof of concept project is described further in *Publication 3: Accuracy and completeness of patient pathways – the benefits of national data linkage in Australia [60].*

1.3. Opportunities and challenges

The development of national linkage infrastructure within Australia presents enormous opportunities for linked research, and has the potential to position Australia as a leader in this area. Further developments in this space have occurred with the release of the Productivity Commission's report *Data Availability and Use* [1] which shows a shift in policy focus from the top of government towards encouraging increased access to data for research.

These developments are not without their challenges, however. The difficulties encountered in the early years of record linkage are essentially the same as those faced today. Despite dramatic advances in computing power, effort is still required to ensure linkage can be completed in a timely manner. An examination of the current record linkage literature reveals that methods for blocking (a record linkage technique whose sole purpose is to speed up operations) continues to be a major research topic [22, 61-63]. Record linkage units continue to develop approaches to optimise their linkage strategy to account for the unreliability of identifying information (see [64-66]). Despite computing advances, manual clerical review still occupies a role in this practice (and remains an expensive and time-consuming process [8, 19]). Issues around privacy remain a major discussion point [67], and have, to some extent, limited the expansion of record linkage [68]. Techniques to reduce privacy risk are another major area of research [13].

This thesis seeks to develop and evaluate improvements in record linkage practice, particularly relating to the challenges of quality and privacy. Before these are discussed, further background is provided into the uses of linked data for research (Chapter 2), and the relevant literature in record linkage methodology is explored (Chapter 3).

Publication 1

Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., Brown, A. P., & Semmens, J. B. (2014). **Technical challenges of providing record linkage services for research**. *BMC medical informatics and decision making, 14*(1), 1.

Contribution:

SR developed the concepts within this paper in collaboration with the other co-authors, and assisted in writing the first draft of the manuscript.

BMC
Medical Informatics & Decision Making

CORRESPONDENCE Open Access

Technical challenges of providing record linkage services for research

James H Boyd[1*], Sean M Randall[1], Anna M Ferrante[1], Jacqueline K Bauer[1], Adrian P Brown[2] and James B Semmens[1]

Abstract

Background: Record linkage techniques are widely used to enable health researchers to gain event based longitudinal information for entire populations. The task of record linkage is increasingly being undertaken by specialised linkage units (SLUs). In addition to the complexity of undertaking probabilistic record linkage, these units face additional technical challenges in providing record linkage 'as a service' for research. The extent of this functionality, and approaches to solving these issues, has had little focus in the record linkage literature. Few, if any, of the record linkage packages or systems currently used by SLUs include the full range of functions required.

Methods: This paper identifies and discusses some of the functions that are required or undertaken by SLUs in the provision of record linkage services. These include managing routine, on-going linkage; storing and handling changing data; handling different linkage scenarios; accommodating ever increasing datasets. Automated linkage processes are one way of ensuring consistency of results and scalability of service.

Results: Alternative solutions to some of these challenges are presented. By maintaining a full history of links, and storing pairwise information, many of the challenges around handling 'open' records, and providing automated managed extractions are solved. A number of these solutions were implemented as part of the development of the National Linkage System (NLS) by the Centre for Data Linkage (part of the Population Health Research Network) in Australia.

Conclusions: The demand for, and complexity of, linkage services is growing. This presents as a challenge to SLUs as they seek to service the varying needs of dozens of research projects annually. Linkage units need to be both flexible and scalable to meet this demand. It is hoped the solutions presented here can help mitigate these difficulties.

Keywords: Medical record linkage, Automatic data processing, Medical informatics computing

Background

Record linkage is the process of bringing together data relating to the same individual from within and between different datasets. When a unique person based identifier exists, this can be achieved by simply merging datasets on the identifier. When this identifier does not exist, some form of data matching or record linkage is required. Often, statistical or probabilistic matching processes are applied to records containing personally identifying information such as name and address.

Record linkage techniques are widely used in public health to enable researchers to gain event based longitudinal information for entire populations. In Australia, research carried out using linked health data has led to numerous health policy changes [1,2]. The success of linkage-based research has led to the development of significant national linkage infrastructure [3]. Comparable record linkage infrastructure exists in few other countries (e.g. England [4], Wales [5], Canada [6], Scotland [7]). The demand for linkage services to support health research, as well as for other forms of human and social research, is increasing [8-10].

There are differing operational models for the provision of record linkage services; however, some elements of the current infrastructure are similar. For example, in Australia and Wales, record linkage is conducted by trusted third parties or specialised linkage units (SLUs). SLUs are usually

* Correspondence: J.Boyd@curtin.edu.au
[1]Centre for Data Linkage, Curtin University, Perth, Western Australia
Full list of author information is available at the end of the article

located external to the data custodians and researchers. This provides an element of separation, which enhances privacy protection [11]. Using specific software, including where appropriate privacy preserving record linkage techniques [12], SLUs engage in high quality data matching. Linkage results (keys) are either returned to the data custodian or forwarded directly to the researcher (depending on the model in use). Once de-identified data has been merged using the linkage keys, analysis of linked data can occur.

The record linkage processes used by SLUs can be quite complex and involve many components e.g. data cleaning and standardisation, deterministic and/or probabilistic linkage, clerical review, etc. Many factors influence the consistency and quality of linkage results [13].

Notwithstanding the complexity of record linkage, SLUs face additional technical challenges in providing linkage 'as a service' for research. The extent of this functionality, and approaches to solving these issues, has had little focus in record linkage literature. Few, if any, of the record linkage packages or systems in use by SLUs today include the full range of functions required of/by these entities.

The purpose of this paper to identify and discuss some of the technical issues associated with the provision of record linkage services, and to propose solutions to these problems. Of particular interest is the array of challenges associated with on-going linkage (i.e. continuous linkage of changing datasets over time). These issues have not been previously addressed in the literature, and it is the aim of this paper to do so.

Methods

The role of SLUs has become more prominent in the research infrastructure landscape and the level and complexity of demands placed on them for linkage services has increased. While there are a variety of techniques available to undertake record linkage such as deterministic rules-based methods, sort and match algorithms [14], and probabilistic techniques [15,16], the tendency for most SLUs has been to implement a probabilistic framework, owing to its robustness, adaptability (particularly in relation to linkage of large datasets – see, for example Clark and Hahn [17]) and high-quality output [18,19]. Probabilistic methods involve sophisticated blocking techniques (to streamline comparisons) and the application of matching methods that incorporate both deterministic and probabilistic comparisons [20-22]. In recent times, there has been extensive work on extending probabilistic approaches and improving efficiency using advances in technology [23,24]. However, beyond the complexity of the linkage process *per se*, there are other technical challenges that present to SLUs. These include the general management of data, handling different linkage scenarios, the management of routine, on-going linkage (and the complexity of storing and handling changing data), the need for automation and the ever present need to accommodate larger sized datasets. In this section we discuss each of these emerging problems.

General management of data

As the number of linkage projects increase, SLUs need robust, efficient methods of managing all forms of data. These include: incoming data from custodians (which need to be maintained in a secure environment, owing to identifying data items and which need to be cleaned and standardised [25] before being used in record linkage); outgoing data (i.e. the linkage keys that are subsequently delivered to others); detailed information about record linkage processes themselves and key decision factors (i.e. linkage strategies, weights, threshold settings, clerical review decisions); linkage results (matched pairs and group membership); and any other value-added information (e.g. geocoding information for addresses).

To ensure robust and reliable linkage operations, the SLUs require close integration between the record linkage software and enterprise level databases. This will help the management of the information resources as the volume of linked data increases.

Handling different linkage scenarios

The linkage requirements of research projects vary. Some research projects require a 'simple' once-off linkage of one or more existing datasets, while others require more intricate linkage of datasets (e.g. genealogical linkage). SLUs need the ability to handle various linkage scenarios including both project based (create and destroy) and ongoing linkage research projects.

'Project based linkage' is arguably the simplest scenario. This is where one or more datasets are required to be linked together for a single research project. These datasets are to be linked to each other, with the links only to be used for a specified research project. Based on the data agreements for the project; the datasets, and the links, often require to be deleted/destroyed after the project has completed.

On-going linkage. As systems, processes and relationships mature, SLUs typically move from a 'project' based approach, where data is linked for each specific research project and then the links are discarded when no longer required, to an on-going approach, where a central core of links is created and maintained over time and re-used for multiple research projects. As new records are added to the system, the links are updated. This approach dramatically reduces effort and improves linkage quality, as the same data are not required to be re-linked over and over with the impact of quality intervention and clerical review is not lost [26]; however, this introduces additional challenges in terms of the volume, speed and quality of matches and the management of associated linkage keys over time is itself complex.

Despite the vast array of record linkage software packages available, most focus on linking files on a 'project' basis, that is, linking a single file to itself (internal linkage) or linking two files to each other at a single instance in time. Currently there are a range of desktop applications that perform this function and although these are usually easy to implement and use, they can struggle to handle medium (>1 million) and large scale (>10 million) linkages [27]. Few, if any, commercial packages exist which have the capacity and functionality to undertake on-going record linkage. As a consequence, these complexities have been resolved in ad hoc ways by individual linkage units.

Alternative approaches to on-going incremental linkage have been developed in recent years, including those outlined by Kendrick [21,28] in his description of Best-link matching. Kendrick's paper expands on the principles outlined by Newcombe [29,30] which describes the factors which could have an effect on the linkage quality, including the likelihood that a record in one file is represented in the matching file.

Other linkage scenarios. There are occasional scenarios where on-going linkage may not be possible, or the most appropriate solution. A SLU needs to understand requirements in both the long and short term, and how it can accommodate both 'project based' and 'on-going' linkage requests, if at all.

Another linkage scenario often dealt with by SLUs is '*bring your own'* linkage. This is where a researcher who has collected information on a study cohort wishes to link this data to another dataset which may or may not already exist in the linkage system. While this researcher's data should link to the required dataset(s), there is no requirement that it should form part of the on-going system.

Challenges associated with on-going linkage

There are several considerations that need to be addressed before implementing an on-going linkage system; these issues typically do not appear in simpler, project based linkage operations. These differences are subtle and are mainly a result of the intricacies of managing data over time. Each of the approaches has their strengths and weaknesses and their applicability or suitability will depend on project requirements.

On-going linkage refers to the process of undertaking *routine, continuous linkage of (changing) datasets over time*. In on-going linkage, previously created links are retained by the system, and added to on the arrival of new records from the same datasets. New records entering the system needed to link to other new records (i.e. internally linked) as well as to existing records that are currently in the system (see Figure 1).

On-going linkage and the management of 'open' records

In project based record linkage, a linkage unit is typically supplied with a series of complete or 'closed' datasets which are required for a research project. These are then linked at a single point in time and the results given to the researcher. In on-going linkage, the necessary datasets are provided to units on a routine and, often, incremental basis. For example, a dataset may be supplied on a monthly basis. This dataset would contain new records for that month, as well as records that were updated during that month. Record received in one month may be amended, or completely removed from the dataset in the next month. An approach to handling new, amended and deleted records is required for on-going linkage.

In order to ensure the integrity of the linkage map and to avoid a re-link of all records, the linkage system should have the ability to detect and handle records which have been amended. This includes records which have had their personal identifying information changed (as these field values may influence matching decisions in earlier iterations of record linkage).

Similarly, the linkage system should have the ability to remove a record from the map. Ideally, this should occur in a way that removes any associations that may have been created by the existence of this record in the system.

Maintaining a linkage map

On-going linkage systems require the maintenance of a central linkage map (a list of each record and the group they belong to). As linkage processes are continuous, the map needs to reflect results *as they occur over time* and

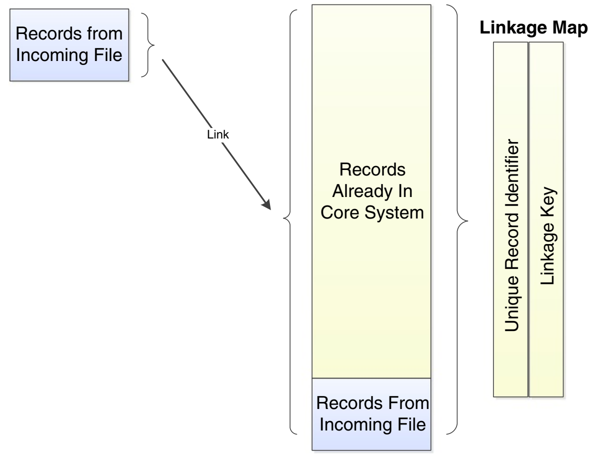

Records from Incoming File
Link
Records Already In Core System
Records From Incoming File
Linkage Map
Unique Record Identifier
Linkage Key

Figure 1 On-going linkage.

for all records in the system, including those that are added or updated on an incremental basis.

Accessing linkage map history

Maintaining a linkage map and its history has utility for researchers, as well as for SLUs. Once researchers receive their data, they may have queries relating to how specific records were linked together. The linkage map is constantly being updated as new records arrive, and as the linkage map may no longer contain these records/links, it may be unclear how these records were brought together. The same problem can occur when a researcher requests a second extraction of their data, (for instance, to receive additional records or content variables). When they receive their second extraction data, they find that the linkage map has changed (as new records have been added or quality fixes have been made) making it difficult to reconcile individual patient histories. For on-going linkage systems, a linkage unit must understand how it will accommodate project requests over time.

Linkage automation

The main goal of adopting on-going linkage is to reduce the amount of time and effort required in conducting a large amount of project linkages, which are routinely re-linking the same data. Taking steps to automate parts of the linkage process fits in naturally with the aim of reducing operator time and effort and increasing scalability.

As on-going linkage systems typically contain a central linkage map which is used in every current and future linkage, the cost of an operator mistake can be very high. Systematic automation and reporting can be useful to ensure and control the quality of linkages over time.

Results

A SLU may employ one of a number of models to ensure that linkage is carried out efficiently and securely while satisfying the linkage needs of the research. Some approaches to automation, linkage scenarios and the creation, management and use of a linkage map are presented below.

Linkage automation

Linkage processes are made up of several discrete steps (as shown in Figure 2), any number of which could be automated. At one end of the spectrum, the grouping process could be automated, with all other processes handled by operators. Upon verifying a file is correct, the operators clean the data and then link the file. When they are satisfied with linkage results, the linkage output is grouped into the linkage map.

Any system containing automation will require a process to ensure tasks are performed in an orderly manner. Looking at the sequence described in Figure 2, for example, a system could be implemented which examined a file to verify it contains the information it was expecting, before cleaning it in a predetermined way, and then linking the file in some predetermined or configurable way. The linkage results could then be added to the linkage map. A fully automated version of such a system would help fulfil the ‘linkage as a service’ model for some SLUs. Linkage services could be further extended so that data providers could connect to a portal to transmit a dataset, which is then automatically linked, with results automatically returned.

There are advantages and disadvantages to automated models of linkage service delivery. Using a fixed approach to cleaning and linking datasets ensures integrity and transparency, and where operators are routinely applying fixed approaches, these could also be added to automated processes. On the other hand, depending on the quality of the data, bespoke approaches to working with individual datasets may improve linkage quality over a one-size-fits-all approach.

Linkage scenarios

Several options exist for handling the different likely linkage scenario requirements. One simple option is to use different linkage systems for different types of linkage scenarios. A SLU may choose to use one set of processes for project-based projects (only), while using an entirely different set of processes/tools for core, on-going linkage. The processes for project linkage may even include manual components.

A more complicated option is to design a single system for all linkage projects but which accommodates differing linkage scenarios for each specific project. Under this option, a linkage project may be configured to be on-going. The associated linkage map would also be ‘on-going’. A linkage project may also be designated to be a hybrid of

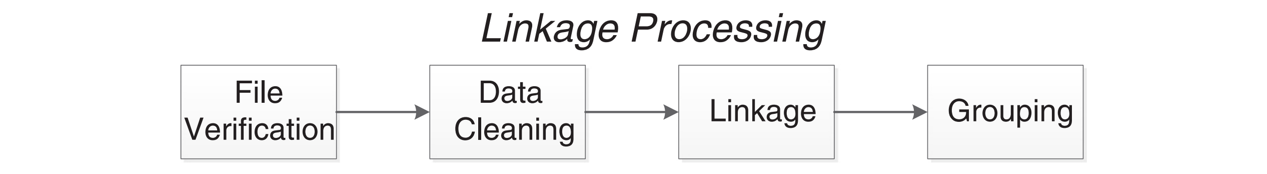


Figure 2 Steps in the linkage process.

projects and on-going linkage, that is, a linkage in which new project datasets are linked to records drawn from an existing, on-going datasets. Linkage results from these project, may, or may not, be added to the on-going linkage map, depending on the requirements of the research project and the likely quality of results.

The most appropriate option will depend, in part, on the number of different linkage scenarios facing SLUs. If requests for separate linkages and linkages to researcher datasets are common, then the first (simpler) option will require a large amount of operator time and resources, defeating the purpose of moving to on-going linkage, while the second may require a large amount of computational resources which may not be feasible.

On-going Linkage

There are several possible methods for conducting on-going linkage and the linkage output will be influenced by a number of factors. One factor is the overlap of people between the files being matched i.e. how many new records have true matches in the existing linked file. Another influence is the size of the existing file, the larger the number of records involved in a probabilistic linkage the greater the likelihood that information will agree 'by chance' across records being compared.

These factors have an influence on the number of records brought together for linkage, the matching strategy and in the post-linkage processes that convert pairs of matched records into groups of records that are stored in a linkage map.

The relationship within and between files and the level of confidence in existing links/relationships are important considerations in the design and optimisation of linkage strategies.

For example, one approach is to link *all* records in the incoming dataset to *all* other records in the system. This method allows pairs to be created describing the relationships between *all* records in the system. With this approach, there are no expectations or assumptions made about how records match against each other or how they group together to become 'sets' of records that belong to the same individual. In terms of linkage strategy, this scenario represents a relatively unconstrained many-to-many linkage. If, however, the linkage task involves linking records to an authoritative record type (i.e. where only one high-quality record per person is known and maintained), then a one-to-one or many-to-one linkage may be more appropriate and there is opportunity to adapt matching strategies to leverage this knowledge [29,30].

A related issue is whether or not to allow merging of groups in the linkage map. A linkage method known as 'best-link matching' [21] makes use of a population spine, which is a set of records already in the system that covers most of the population, and has been linked to a high standard. In this method, incoming records are unable to join together two groups already existing in the system– instead the 'best link' is chosen, and the incoming record is added to this group (Figure 3, Option 1).

This method uses underlying knowledge of the quality of the population spine to make decisions about future linkage results. Most SLUs accept that a small percentage of matches will be incorrect. In the situation where one of these matches merges two groups, the error is compounded and all records within these two groups are now incorrectly linked together[a].

An alternate approach is to allow the merging of groups to occur. This method does not rely on the existence of a high quality reference dataset (spine). For this reason this method may be useful in a much greater range of circumstances.

There is an additional advantage to choosing strategies which allow merging of groups and which use all records in linkage. The advantage of this approach (and this approach only), is that the order of the incoming records does not affect system groupings. It is intuitive that this should be the case, as in practice the order of received records is typically highly dependent on contractual arrangements and other arbitrary preparations, which should not have an effect on the groups made by the system.

Managing and accessing a changing linkage map

In on-going linkage, the linkage map is constantly changing and there may be requests from researchers to access results from previous linkages. There are several ways in which a SLUs can manage changing linkage maps and accommodate requests for past information. One solution is to take snapshots of the linkage system at the point of extraction for all research projects. This allows researchers access to the data and linkage map at the time of extraction and will solve the majority of the researchers queries, although the system would not be able to determine exactly why things have changed. While multiple snapshots of the system would take up a large amount of space, these do not necessarily need to be stored on on-going infrastructure, and could be moved elsewhere until required.

An alternative solution is to have a linkage map which stores the full history of groups, recording details of when additional records entered or left specific groups. This allows full understanding of how groups of records came together, as well as giving the ability to 'roll-back' to a point in time when an extraction for a researcher occurred (see Figure 4). Storing the full history of groups will likely take up more space in the linkage map; however, it provides greater flexibility in the extraction process and changes to groups are fully documented.

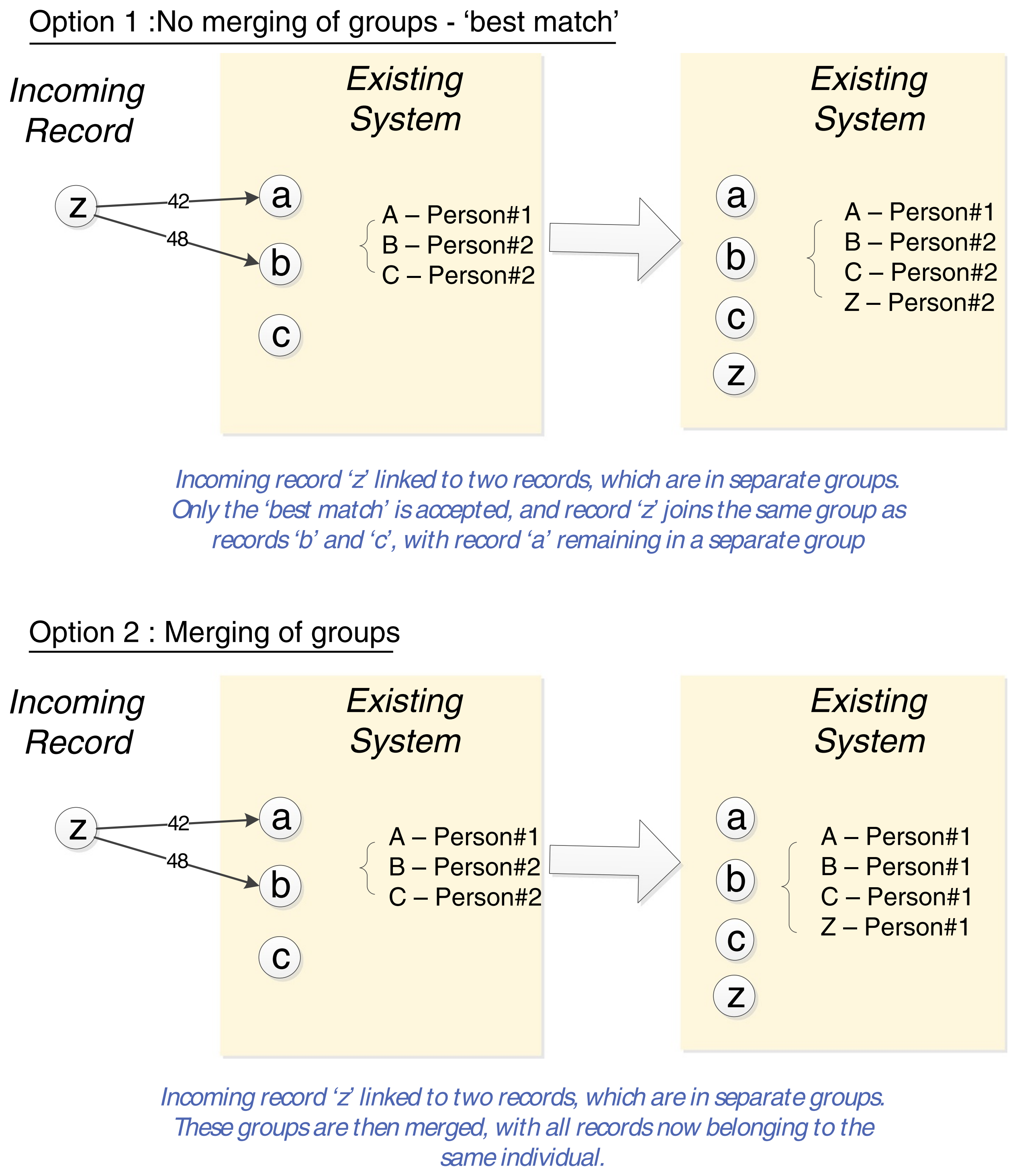

Option 1 :No merging of groups - 'best match'
Incoming Record
Existing System
z
42
48
a
b
c
A – Person#1
B – Person#2
C – Person#2
Existing System
a
b
c
z
A – Person#1
B – Person#2
C – Person#2
Z – Person#2
Incoming record 'z' linked to two records, which are in separate groups. Only the 'best match' is accepted, and record 'z' joins the same group as records 'b' and 'c', with record 'a' remaining in a separate group
Option 2 : Merging of groups
Incoming Record
Existing System
z
42
48
a
b
c
A – Person#1
B – Person#2
C – Person#2
Existing System
a
b
c
z
A – Person#1
B – Person#1
C – Person#1
Z – Person#1
Incoming record 'z' linked to two records, which are in separate groups. These groups are then merged, with all records now belonging to the same individual.

Figure 3 Methods for on-going linkage.

Managing deleted, amended and 'open' records

Deleted records

One option for managing deleted records is simply to remove them from the groups they are currently part of. The danger with this method is that the deleted record may have erroneously brought together two groups of records, which may now stay together indefinitely. A better approach is to unwind these groups by utilising the

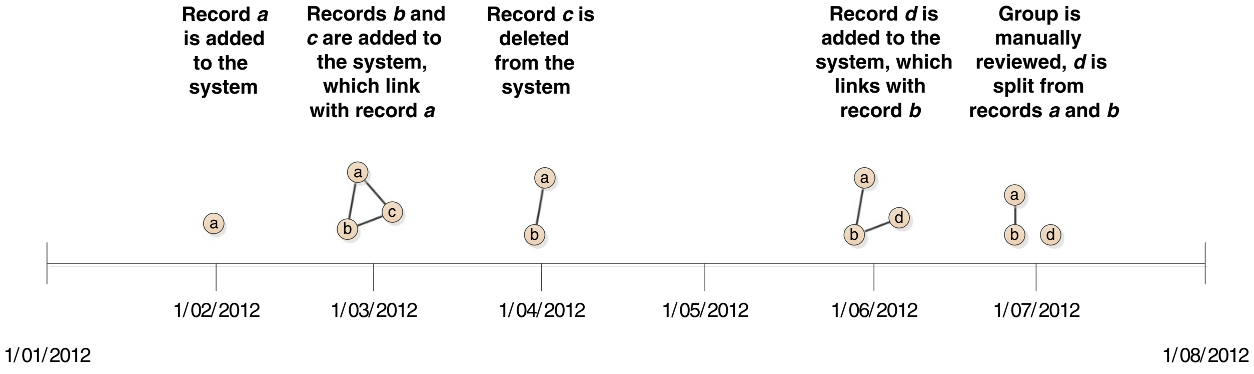

Record *a* is added to the system
Records *b* and *c* are added to the system, which link with record *a*
Record *c* is deleted from the system
Record *d* is added to the system, which links with record *b*
Group is manually reviewed, *d* is split from records *a* and *b*
a
a
b
c
a
b
a
b
d
a
b
d
1/01/2012
1/02/2012
1/03/2012
1/04/2012
1/05/2012
1/06/2012
1/07/2012
1/08/2012

Figure 4 The full history of stored groups including the reasons for changes.

matching pair information used in creating these groups to discover how these groups would have looked had this deleted record not entered the system (Figure 5).

Amended records

There are several options available to manage amended records. One option is simply to amend the details stored in the database, without changing the system groupings. However these amended details may mean this record should belong to a different group, and that these links are actually in error.

An alternative option is to treat the amended record as a new record. In order to ensure the integrity of the linkage map, one must also identify and re-link any records that previously match to the record. This will ensure the new version is linked to the appropriate records.

By using pair information during deletion, and re-linking amended records, we can ensure the linkage map

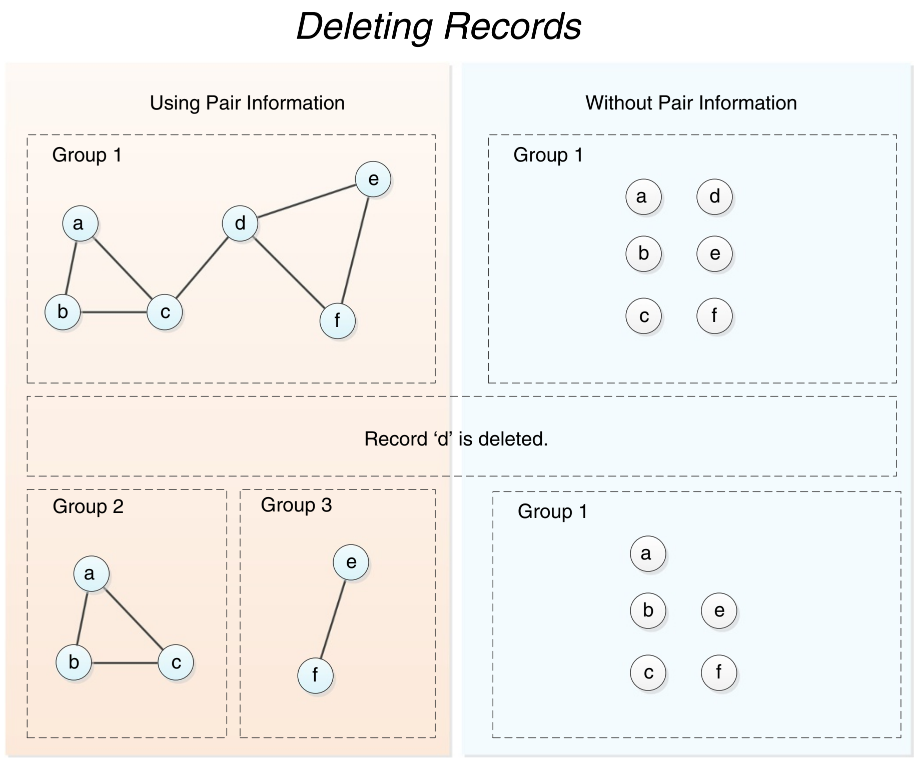


Figure 5 Methods for deleting records.

looks the same as if the deleted records and previous version of the amended records had never entered the system.

'Open' records

Linkage systems that can handle deleted and amended records are better placed to accommodate the linkage of 'open' records. 'Open' records are those records where creation and end times vary and where the content of data may change between those dates. Many data providers only work with 'closed' records, which they can guarantee will not change. This process involves extensive validation and cleaning of the data before the file can be closed. This process is time consuming but ensures no changes to the linkage map once the file has been added. Some collection systems have 'open' records which can be amended over time. The advantage of 'open' files is that they can be updated to reflect amendments to records or deletions.

Discussion

SLUs must service a range of record linkage needs from the research community. They must be able to deal with a range of linkage scenarios, from (simple) project linkage based approaches to complex on-going linkage. On-going linkage requires consideration of a number of additional time-sensitive issues which do not affect project based linkages. Despite the complexity, the advantages of moving to a more automated, efficient and sustainable way of conducting linkage far outweigh the intricacies of doing so. Table 1 summarises these key operational features of a linkage system and options available.

Several themes run throughout the issues presented in this paper. One is the trade-off between automation and bespoke approaches. Bespoke approaches will always be more flexible, but will always suffer from issues of transparency, maintainability and replicability. A second theme is the focus on issues and processes that complement and support the specialised activities of record linkage units. As presented in this paper, there are a number of key technical issues which must be understood and overcome in order for SLUs to deliver efficient record linkage 'services' for researchers.

There are several areas of further research required. To our knowledge, none of the options presented in this paper have been empirically compared against each other. However the employment of one option over another depends (typically) on assumptions about linkage quality, a measurable trait. If empirical research investigated the effect on linkage quality of several of these options over time given different datasets and other parameters, linkage units would be better equipped to decide on the most appropriate option for their systems.

A second area of research is related to the benefit of bespoke processes over automated processes. While it is assumed that automatic processes will likely produce lower quality results, the actual degradation in quality is not known. Research which tests and quantifies these effects is warranted. Until we know the true effect that automation has on linkage quality (if any), linkage units cannot make an informed decision about the benefit of this move.

Conclusion

The process of conducting numerous linkages on a large scale is both complex and resource intensive. Linkage systems need to be both flexible and scalable to meet the future demands of enterprise-level record linkage. It is hoped the solutions presented here help reduce these difficulties.

Table 1 Summary of issues and options for on-going linkage

Operational feature	Options
On-going linkage	- Link to most recent record in group vs. link to all records
	- Best-link matching vs. merging groups
Linkage automation	- Spectrum from fully automated to only the grouping process automated
Links stored	- No history stored
	- Snapshots stored
	- Full history stored within linkage map
Handling different linkage scenarios	- Only on-going linkage
	- Manual processes for project based linkage
	- Access to on-going linkage system used for project based linkage
	- Build system which can handle multiple scenarios
Amended and deleted records	- No handling of amended and deleted records
	- Amended records: Changing personal identifiers only vs deleting and re-linking
	- Deleted records: Simple removal, or using pair information to reconstitute groups

Endnote

[a]In this method false negatives found in the originating dataset used for the population spine will never be brought together no matter what additional information is found in other datasets. Additional records can provide new information which makes it clear that two records previously existing within the system actually belong to the same person. In these situations, ‘best-link matching’ will not be able to use this information to improve quality.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Initial design and conception provided by JHB, AF and JS. Further technical design provided by AB, JKB and SR. First draft of manuscript provided by SR; subsequently edited significantly by JHB, AF and JS. All authors read and approved the final manuscript.

Acknowledgements

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy’s Population Health Research Network. The authors would like to thank the reviewers for their invaluable comments.

Funding statement

This research received no specific funding.

Author details

[1]Centre for Data Linkage, Curtin University, Perth, Western Australia.
[2]The Birchman Group, Perth, Western Australia.

Received: 26 June 2013 Accepted: 25 March 2014
Published: 31 March 2014

References

1. Brook EL, Rosman DL, Holman CDAJ: **Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System.** *Aust N Z J Public Health* 2008, **32**(1):19–23.
2. Hall SE, Holman CDAJ, Finn J, Semmens JB: **Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records.** *Int J Qual Health Care* 2005, **17**:375–381.
3. Boyd JH, Ferrante AM, O’Keefe CM, Bass AJ, Randall SM, Semmens JB: **Data linkage infrastructure for cross-jurisdictional health-related research in Australia.** *BMC Health Serv Res* 2012, **12**(1):480.
4. Gill LE: **OX-LINK: the oxford medical record linkage system.** In *Record Linkage Techniques – 1997.* Edited by Alvey W, Jamerson B. Washington DC: National Academy Press; 1999:15–33.
5. Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, Brooke CJ, Thompson S, Bodger O, Couch T, Leake K: **The SAIL Databank: building a national architecture for e-health research and evaluation.** *BMC Health Services Research* 2009, **9**(1):157.
6. Roos LL, Nicol JP: **A research registry: uses, development, and accuracy.** *J Clin Epidemiol* 1999, **52**(1):39–47.
7. Kendrick S, Clarke J: **The Scottish record linkage system.** *Health Bull* 1993, **51**(2):72.
8. OECD: *Strengthening Health Information Infrastructure for Health Care Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges.* OECD Publishing; 2013.
9. Ferrante A: **The use of data-linkage methods in criminal justice research: a commentary on progress, problems and future possibilities.** *Curr Issues Crim Justice* 2009, **20**(3):378–392.
10. Jutte DP, Roos LL, Brownell MD: **Administrative record linkage as a tool for public health research.** *Annu Rev Public Health* 2011, **32:**91–108.
11. Kelman C, Bass J, Holman D: **Research use of linked health data: a best practice protocol.** *Aust N Z J Public Health* 2002, **26**(3):251–255.
12. Schnell R, Schnell T, Bachteler J, Reiher: **Privacy-preserving record linkage using Bloom filters.** *BMC Med Inform Decis Mak* 2009, **9**(1):41.
13. Roos L, Wajda A: **Record linkage strategies. Part I: estimating information and evaluating approaches.** *Methods Inf Med* 1991, **30**(2):117.
14. Hernández MA, Stolfo SJ: **Real-world data is dirty: data cleansing and the merge/purge problem.** *Data Min Knowl Discov* 1998, **2**(1):9–37.
15. Fellegi I, Sunter A: **A theory for record linkage.** *J Am Stat Assoc* 1969, **64:**1183–1210.
16. Newcombe H, Kennedy J: **Record linkage: making maximum use of the discriminating power of identifying information.** *Commun ACM* 1962, **5**(11):563–566.
17. Clark DE, Hahn DR: **Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry.** *Proc Annu Symp Comput Appl Med Care* 1995, **1995:**397–401.
18. Pinder R, Chong N: **Record linkage for registries: current approaches and innovative applications.** In *Presentation to the North American Association of Central Cancer Registries Informatics Workshop. Toronto, Canada*; 2002.
19. Gomatam S, Carter R, Ariet M, Mitchell G: **An empirical comparison of record linkage procedures.** *Stat Med* 2002, **21:**1485–1496.
20. Roos LL, Wajda A, Nicol JP: **The art and science of record linkage: methods that work with few identifiers.** *Comput Biomed Med* 1986, **16**(1):45–47.
21. Kendrick S, Douglas M, Gardner D, Hucker D: **Best-link matching of Scottish health data sets.** *Methods Inf Med* 1998, **37**(1):64.
22. Winkler WE: **Advanced methods for record linkage.** In *Statistical Research Report.* Washington D C: U S Bureau of the Census, Statistical Research Division; 1994.
23. Winkler WE: In *Using the EM algorithm for weight computation in the Fellegi-Sunter Model of record linkage.* Edited by Census UBot. Washington DC; 2000:12.
24. Herzog TH, Scheuren F, Winkler WE: **Record linkage.** In *Wires Computational Statistics.* New York: John Wiley Sons; 2010:9.
25. Randall SM, Ferrante AM, Boyd JH, Semmens JB: **The effect of data cleaning on record linkage quality.** *BMC Med Inform Decis Mak* 2013, **13**(1):64.
26. Rosman D, Garfield C, Fuller S, Stoney A, Owen T, Gawthorne G: **Measuring data and link quality in a dynamic multi-set linkage system.** In *Symposium on Health Data Linkage.* Sydney, NSW; 2002. http://www.publichealth.gov.au/pdf/reports_papers/symposium_procdngs_2003/rosman_a.pdf.
27. Ferrante A, Boyd J: *Data Linkage Software Evaluation: A First Report (Part I).* Perth: Curtin University; 2010.
28. Kendrick SW, McIlroy R: *One Pass Linkage: The Rapid Creation of Patient-based Data. Proceedings of Healthcare Computing1996.* Weybridge, Surrey: British Journal of Healthcare Computing Books; 1996.
29. Newcombe HB: *Handbook for Record Linkage: Methods for Health and Statistical Studies, Administration and Business.* New York: Oxford University Press; 1988.
30. Newcombe H: **Age-related bias in probabilistic death searches Due to neglect of the “Prior Likelihoods”.** *Comput Biomed Res* 1995, **28**(2):87–99.

doi:10.1186/1472-6947-14-23
Cite this article as: Boyd *et al.*: Technical challenges of providing record linkage services for research. *BMC Medical Informatics and Decision Making* 2014 **14**:23.

Publication 2

Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., **Randall, S. M.**, & Semmens, J. B. (2012). **Data linkage infrastructure for cross-jurisdictional health-related research in Australia.** *BMC health services research, 12*(1), 1

Contribution:

SR supported the development of this paper and made contributions to the final version of the manuscript.

BMC
Health Services Research

CORRESPONDENCE Open Access

Data linkage infrastructure for cross-jurisdictional health-related research in Australia

James H Boyd[1*], Anna M Ferrante[1], Christine M O'Keefe[2], Alfred J Bass[3], Sean M Randall[1] and James B Semmens[1]

Abstract

Background: The Centre for Data Linkage (CDL) has been established to enable national and cross-jurisdictional health-related research in Australia. It has been funded through the Population Health Research Network (PHRN), a national initiative established under the National Collaborative Research Infrastructure Strategy (NCRIS). This paper describes the development of the processes and methodology required to create cross-jurisdictional research infrastructure and enable aggregation of State and Territory linkages into a single linkage "map".

Methods: The CDL has implemented a linkage model which incorporates best practice in data linkage and adheres to data integration principles set down by the Australian Government. Working closely with data custodians and State-based data linkage facilities, the CDL has designed and implemented a linkage system to enable research at national or cross-jurisdictional level. A secure operational environment has also been established with strong governance arrangements to maximise privacy and the confidentiality of data.

Results: The development and implementation of a cross-jurisdictional linkage model overcomes a number of challenges associated with the federated nature of health data collections in Australia. The infrastructure expands Australia's data linkage capability and provides opportunities for population-level research. The CDL linkage model, infrastructure architecture and governance arrangements are presented. The quality and capability of the new infrastructure is demonstrated through the conduct of data linkage for the first PHRN Proof of Concept Collaboration project, where more than 25 million records were successfully linked to a very high quality.

Conclusions: This infrastructure provides researchers and policy-makers with the ability to undertake linkage-based research that extends across jurisdictional boundaries. It represents an advance in Australia's national data linkage capabilities and sets the scene for stronger government-research collaboration.

Keywords: Data linkage, Infrastructure, Population, Health, Research

Background

Benefits of data linkage to research, policy making and service delivery

Administrative datasets constitute a significant information resource for government and are used to manage, monitor, assess and review a range of service areas. They are also used in research to provide insight into significant health issues, to support health policy development and improve clinical practice and service delivery. Additional value can be obtained from these administrative collections through data linkage. This process allows data from different sources, including disease registers and clinical datasets, to be brought together to provide richer information. The benefits of linked data include reduced data collection costs and more detailed and extensive analysis [1-6].

Data linkage infrastructure developments

Despite recognition of the value of data linkage by government and the research community, dedicated infrastructure to sustain and support data linkage activity is limited. Data linkage "systems" or "facilities" exist in only a handful of countries including Canada [7], England (Oxford) [8], Scotland [9], Australia [10] and most recently in Wales through the development of the SAIL system [11]. These production-level systems undertake linkage on a routine

* Correspondence: j.boyd@curtin.edu.au
[1]Curtin University, Perth, Western Australia
Full list of author information is available at the end of the article

basis, servicing the statistical and research needs of both government and University researchers.

In Australia, purpose-built data linkage infrastructure was first established in 1995 in Western Australia. The Western Australia Data Linkage System (WADLS) emerged from a collaboration between the University of Western Australia's School of Population Health and the Western Australia (WA) Department of Health. WADLS comprises a complex probabilistic data matching system to create, store, update and retrieve links between over 40 population-based administrative and research health data collections in WA [12]. Following the success of the WADLS and in recognition of the power of the resulting linked research data, the Centre for Health Record Linkage (CHeReL) was established in 2006 in New South Wales (NSW) to undertake data linkage for NSW and the Australian Capital Territory [13]. Hosted by the NSW Cancer Institute, CHeReL is a joint venture between eight institutions. It has developed quickly to incorporate the routine linkage of a number of strategic, core datasets.

PHRN initiative

Further investment in Australia's data linkage capability occurred in 2006 when the Australian government allocated $20 million to further develop data linkage infrastructure under the National Collaborative Research Infrastructure Strategy (NCRIS). State and Territory governments and academic partners invested a further $32 million to support the capability. The initiative, known as the Population Health Research Network (PHRN), included the establishment of data linkage units in all other Australian States, the formation of the Centre for Data Linkage (CDL) for national or cross-jurisdictional linkage, the development of a secure remote access laboratory for researchers, and a data delivery system for the secure electronic transfer of data between PHRN participants and relevant stakeholders. The purpose of the PHRN is to "provide researchers in Australia with the capability to link de-identified data from a diverse and rich range of health datasets, across jurisdictions and sectors, to carry out nationally and internationally significant population-level research, to improve health and wellbeing and enhance the effectiveness and efficiency of health services" [14].

A core component of the PHRN infrastructure has been the development of national or "cross-jurisdictional" linkage capability i.e. the ability to link data from more than one State or Territory. Given the federated nature of health care service delivery in Australia (i.e. some services are delivered and administered at State level, while others are delivered and administered at a national or "Commonwealth" level), cross-jurisdictional linkage is an essential component of national infrastructure. Without cross-jurisdictional data linkage capabilities, research aimed at national level or targeting issues of common interest (e.g. health service use along border areas) cannot be undertaken. The remainder of this paper describes the development of the processes and data linkage methodology required to create a cross-jurisdictional research infrastructure and the aggregation of State and Territory linkages into a single system.

Methods

Under the PHRN initiative, the CDL was tasked with "establishing a secure and efficient data linkage system to facilitate linkage between jurisdictional datasets, and between these datasets and research datasets using demographic data" [14]. To fulfil this function, the CDL engaged in the:

i) Development of a cross-jurisdictional operational model
ii) Specification and implementation of a secure IT environment including linkage software; and
iii) Development and adoption of strong governance arrangements

CDL operational model development

The operations and infrastructure in the CDL build on the models created in both WADLS and CHeReL. The Cross-Jurisdictional Operational Model was developed in wide and open consultation with PHRN members and related stakeholders [15]. The Model incorporates a separated and layered linkage approach where State/Territory linkages are conducted by individual State-based or "jurisdictional" linkage units, while cross-jurisdictional or "national" linkages are conducted by the CDL (see Figure 1).

This layered model maximises the skills and experience in data linkage across Australia and builds on the success of well established data linkage units in WA and NSW/ACT. It involves a multi-tier operating structure with standardised governance arrangements which are responsive to researchers needs. The state/territory data linkage units have had a major influenced on the development of the model and the CDL has benefited from working with state/territory data linkage units to understand the data, the technologies and researcher needs. The layered model also allows efficient control over aspects such as skill development, resource utilisation, operational efficiency and the application of standards across data linkage units.

A best practice 'separation' principle operates in the Model at both State (or "jurisdictional") and CDL levels [16]. Under this principle, the process of data linkage (and the data items used in linkage activity) is kept separate from the processes that extract and deliver content

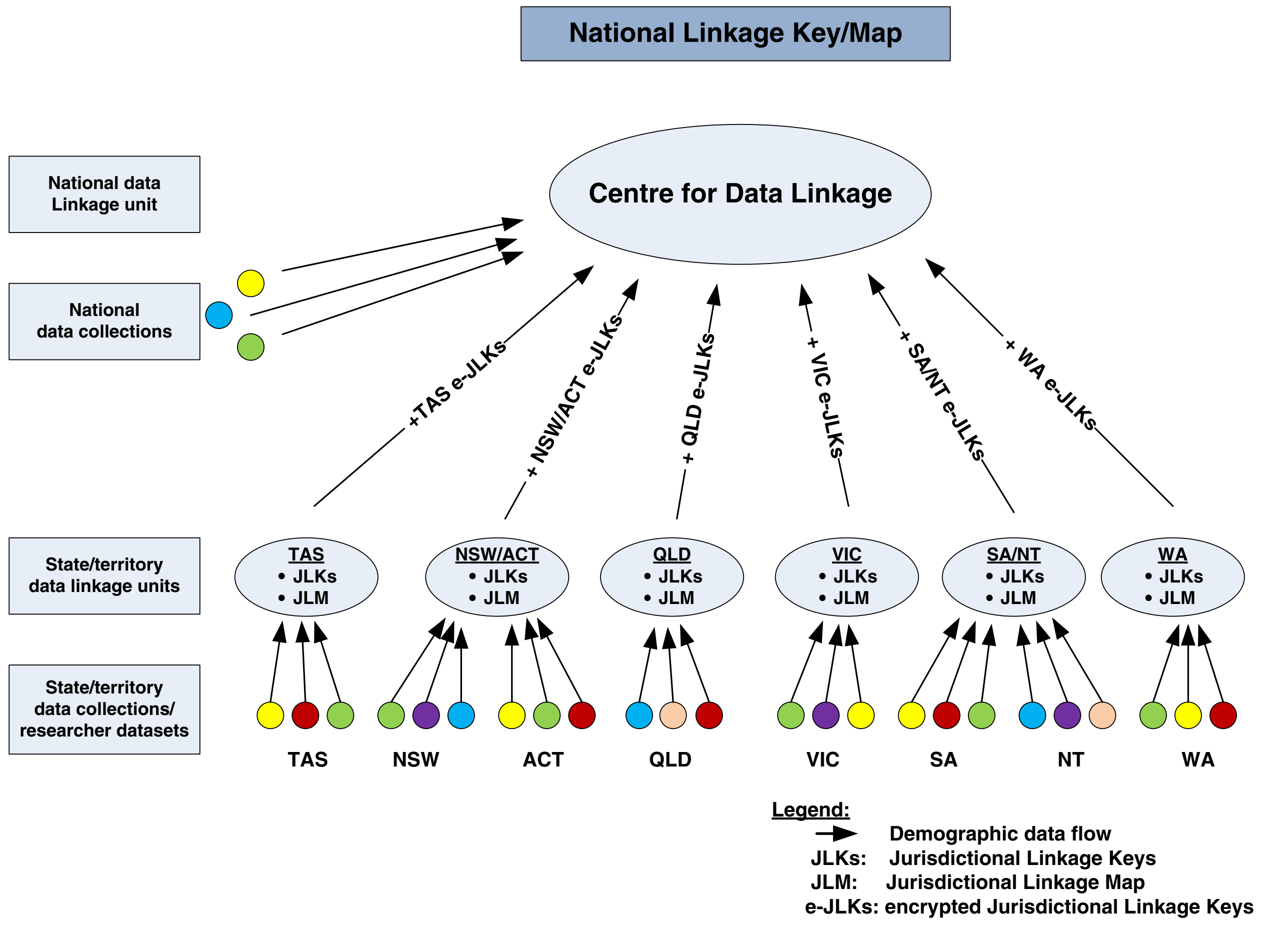


Figure 1 Layered Linkage Model of the PHRN.

or clinical data for researchers. Data flows for cross-jurisdictional linkage comprise three distinct phases:

- Flow of data for linkage
- Provision of project specific linkage keys
- Extraction of research data

Phase One of the data flow model is about **the linkage process**. The data used for linkage involves only a limited set of variables, typically demographic data (e.g. name, date of birth, address, date of event). This information is used for linkage purposes only. A Data Custodian provides demographic data and related record identifiers to the Jurisdictional Data Linkage Unit. The Jurisdictional Linkage Unit uses this data to undertake state-based linkages for state-based research projects. For cross-jurisdictional projects, the local Linkage Unit provides the demographic data and encrypted record identifiers to the CDL. The CDL uses this information to link data across multiple jurisdictions.

An important element of the Cross-Jurisdictional Model is the creation and maintenance of a National Linkage Map [17]. Following the linkage process, the CDL assigns the same reference key – a National Linkage Key (NLK) - to each record that is considered to belong to the same person. The reference between the Unique Record Identifier (RecIDs) of each record and the NLK creates the national linkage map (i.e. a direct list showing the national linkage key corresponding to each unique record identifier). Allocation of the NLKs allows the system to group records within the National Linkage Map to show which sets of entries are considered to refer to the same person.

Each NLK only has value within the context of the National Linkage Map, which associates them with pointers to health records. The Unique Record Identifiers contained in the Map are encrypted and each is used as a pointer to the information held by data providers. It is important to note that the National Linkage Map does not contain any demographic or content variables. When extracted, information from the National Linkage Map are masked and then encrypted before being supplied to Data Custodians for approved research projects. Phase Two of the process is the **provision of project-**

specific linkage keys which enables research datasets to be extracted and merged anonymously by researchers. For each cross-jurisdictional project, the CDL returns to the local Jurisdictional Linkage Unit a file with the record identifiers and project-specific linkage keys. Each project is issued with a unique set of project-specific linkage keys. The local Linkage Unit passes the project-specific key and record identifiers to the Data Custodian who then proceeds to the final phase of the process (data extraction).

Phase Three, **extraction of research data** for approved projects, takes place only after Phase One and Phase Two have been completed. For each cross-jurisdictional research project, content data is extracted by the Data Custodian. It consists of project-specific linkage keys and only those variables which the researcher has been authorized to access. The dataset does not contain any identifying data items (e.g. name). The linkage keys in the dataset are project-specific so that researchers cannot collude and bring together data from different projects. Once the researcher is provided with data from all relevant Data Custodians, records can be merged using the project-specific linkage key and then used in analyses.

As Figure 2 shows, the Data Custodian is an integral part of all steps of the process and directly controls access to their data. This Model does not involve a central data repository which means that custodians only release data on a project by project basis. The CDL does not hold clinical or content data, but links the demographic data that has been separated from the remainder of each dataset to create ‘linkage keys’. Clinical or service information is not needed by the CDL and is not provided to it and the researcher receives only that part of the record that they have approval to see (without any demographic or identifying information).

With the model separating the linkage and research data and functions, access to reliable metadata during the linkage and analytical part of each cross jurisdictional research project is important. In Australia the METEoR system is one such metadata repository that provides a single-source dataset of definitions (including those administrative in nature) at a national level. This will be a useful resource to align the definitions across jurisdictional datasets.

Secure IT environment

To implement the Operational Model, the IT infrastructure arrangements for CDL had to provide a secure controlled environment for working with name-identified data. Understanding the sensitive nature of identifying information assets, the CDL designed its operations to accommodate datasets from State and Commonwealth organisations whilst applying the highest level of security. As well as ensuring that identifying demographic information was handled separately from any content or clinical data as part of its data flows, the CDL established a secure IT infrastructure to protect these information assets throughout the process.

A secure stand-alone network (the CDL stand-alone network) was designed in consultation with the PHRN to enable the storage and processing of demographic data

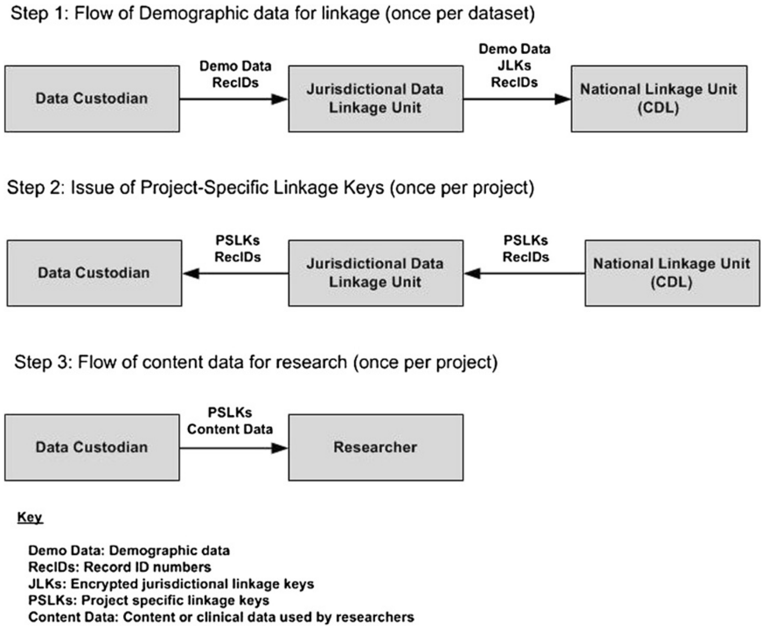

Step 1: Flow of Demographic data for linkage (once per dataset)
Data Custodian
Demo Data
RecIDs
Jurisdictional Data Linkage Unit
Demo Data
JLKs
RecIDs
National Linkage Unit (CDL)
Step 2: Issue of Project-Specific Linkage Keys (once per project)
Data Custodian
PSLKs
RecIDs
Jurisdictional Data Linkage Unit
PSLKs
RecIDs
National Linkage Unit (CDL)
Step 3: Flow of content data for research (once per project)
Data Custodian
PSLKs
Content Data
Researcher
Key
Demo Data: Demographic data
RecIDs: Record ID numbers
JLKs: Encrypted jurisdictional linkage keys
PSLKs: Project specific linkage keys
Content Data: Content or clinical data used by researchers

Figure 2 Cross-jurisdictional data flows.

received from the jurisdictional linkage units, researchers and other sources. The Australian Department of Defence publication ACSI 33 Australian Government ICT Security Manual (ISM) was used as a guideline for identifying risks and controls when considering requirements and determining CDL security measures. The ISO/IEC 17799:2005 Information Technology – Security Techniques – Code of Practice for Information Security Management was also consulted in developing the CDL IT solution and security plan. As Figure 3 demonstrates, the CDL stand-alone network is physically separate from all other networks. The environment was later subjected to an independent, external security audit.

Independent audit

The objectives of the independent audit were to review the CDL secure IT environment, and identify and describe the controls to ensure that they were being applied in compliance with the standards and processes identified by the PHRN stakeholders. The audit included a full review of the configuration, operations, and usage of the CDL infrastructure.

Among other things, the audit report provided an assessment of how the infrastructure was configured and used relative to the standards identified by the PHRN stakeholders and recommended changes to configuration and usage.

Governance

A major challenge for all members of the PHRN has been to ensure that the collection, use and disclosure of personal information comply with applicable information privacy legislation. Compliance with legal requirements relating to privacy is essential but it is only one dimension of good governance. Equally important is the development of a strong culture of understanding and support for privacy goals and governance best practice.

Among the governance structures instituted by the PHRN are a Management Council overseeing the implementation of the national data linkage program, with sub-committees which provide advice and direction to Management Council members. These sub-committees include an Ethics, Privacy, and Consumer Engagement Advisory Group, an Operations Committee (providing technical advice) an Access Committee (providing advice on access, accreditation and eligibility); a Data Transfer Working Group and Proof of Concept Reference Group. Additional governance features of the PHRN include a

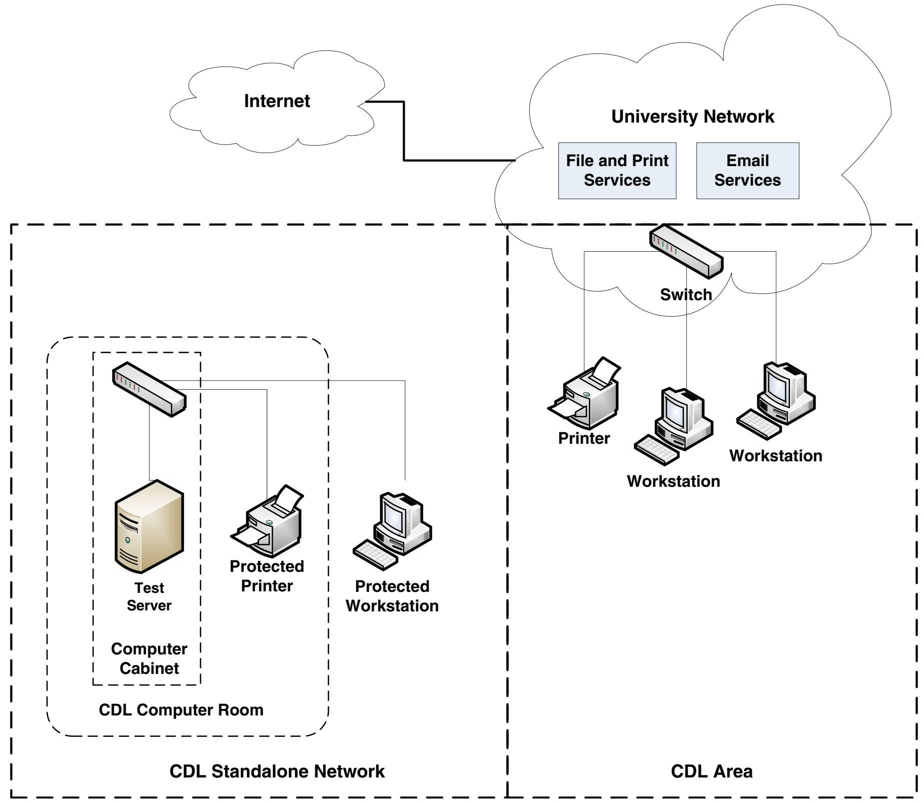


Figure 3 CDL IT infrastructure configuration. * For Proof of Concept, computing capabilities were enhanced; however, the basic configuration of IT infrastructure was unchanged. Additional computing resources included two new servers and UPS.

strict reporting regime; a Privacy framework; an Information Governance framework; rigorous approvals processes for each research project; binding agreements related to data release, date confidentiality and security and Network-wide policies and guidelines.

Software evaluation

A need to identify accurate, reliable, load-bearing (i.e. production capability) record linkage software was recognised in the very early stages of development. As a consequence, the CDL embarked on an evaluation of ten data linkage software packages to assess their suitability for inclusion in a large scale automated production environment [18,19]. The evaluation identified three potential candidate packages. These products were shortlisted for further testing during the Proof of Concept phase (POC; see below).

PHRN proof of concept linkages

The primary aim of the PHRN Proof of Concept projects is to demonstrate the capability of the PHRN infrastructure to answer research questions of national importance, by conducting inter-state linkages [14]. The first PHRN Proof of Concept project examined in-hospital mortality and investigated issues of hospital safety and quality using inpatient and mortality information.

Initial data was provided to the CDL from NSW and WA. This comprised more than 25 million hospital and mortality records over a ten year period. Consistent with the Cross-Jurisdictional Model, data flows and linkage activity included the following:

- Transfer of hospital and mortality demographic information and jurisdictional linkage keys from custodians and linkage units in NSW and WA to the CDL
- Linkage of this data to create a national map
- Creation of project-specific linkage keys based on this map
- Transfer project-specific linkage keys back to the jurisdictions
- Transfer of the necessary clinical data from the jurisdictional custodians to the researcher

Results and discussion

The CDL Cross-Jurisdictional Model was endorsed by the PHRN Management Council in 2010 [20]. A development and implementation programme based on that Model subsequently commenced (and is still on-going). The development programme includes the design and implementation of a large-scale automated production linkage system in which a national linkage map can be created and maintained over time as new datasets and updates to datasets become available.

Strengths and weaknesses of the model

The Cross-Jurisdictional Model has a number of design strengths. Firstly, it implements the best practice separation model [16] to protect the privacy of individuals. Secondly, it adopts a “minimum data” principle in which participants are provided only with the minimum amount of information required to conduct their designated activity. Both of these elements are consistent with Australian government principles for data integration [21]. The Jurisdictional Linkage Units and encrypted versions of their jurisdictional linkage keys are integral to the process. They ensure that high quality linkages at both state and national level are maintained and that resources are used efficiently. The independence of Jurisdictional Linkage Units is also maintained under this Model, as is the proximal relationship between these Units and local data custodians. Finally, the Cross-Jurisdictional Model is designed to be extensible – datasets and/or linkage units can be added with minimum impact on the overall system.

Although the Model has been designed to maximise the protection of privacy, the additional data flows also introduce some operational restrictions. The obvious limitation is around the coordination of numerous “separated” elements before different datasets can be joined up. This process can be complex and requires careful consideration to avoid bottlenecks in the system. There are other limitations to the Model. For example, there is no flexibility in operations – roles of participants are defined from the start. Data flows are also likely to be slow and highly dependent on the capabilities and resourcing of Data Custodians. Processes may be difficult to speed up or streamline. System auditing is also more difficult under a “separated” Model, as it is difficult to trace the history of linked analytical data without good coordination and oversight.

This model was agreed to after extended consultation with the rest of the network. A consultation paper was presented to PHRN participants outlining proposed models and asking for feedback regarding particular options. The model was chosen based on a desire to find consensus amongst participants. Alternative models were proposed, including the CDL receiving data directly from state Data Custodians. Receiving data from linkage units allowed the CDL to leverage off the existing relationship between the data custodians and linkage units, and to utilise the jurisdictional linkage keys for quality assurance purposes.

Operational governance and IT

The CDL has established a development programme which involves constructing effective matching methodologies around the agreed operational model. In addition to developing and demonstrating technical linkage capabilities, governance arrangements at the CDL were further developed and refined. The CDL has developed specific governance provisions around security and operations, risk management and privacy (including Privacy Impact Assessment). Ethics approval has been granted to operate the CDL cross-jurisdictional data linkage infrastructure.

A secure IT environment was established to meet the security standards developed as part of the PHRN Information Governance Framework for cross-jurisdictional data linkage. The environment was later subjected to an independent, external security audit as part of the threat and risk assessment process.

Overall the audit concluded that the CDL environment and systems were being managed in an efficient and reliable manner. Although no major deficiencies were observed, the report provided non-essential recommendations. All recommendations were addressed successfully. The independent audit review process has been included in the CDL Governance Plans which means that other audits will be required in the future if there are significant changes to the secure IT environment.

Software evaluation

The software evaluation was successful in identifying appropriate software for production linkage. The software evaluation also resulted in the development of a unique, sharable methodology for data linkage software evaluation. The methodology incorporates the use of synthetic data and is both transparent and transportable [22]. The knowledge and expertise developed through the evaluation was shared with the wider PHRN to assist their developments.

PHRN proof of concept linkages

The cross-jurisdictional data linkage capabilities of the CDL have been demonstrated through involvement in the PHRN Proof of Concept Collaboration projects. Using its data linkage capabilities, the CDL linked both NSW and WA data as new and compared these results to those achieved by the WA Data Linkage Branch (WADLB) and the NSW CHeReL. The jurisdictional linkage keys supplied by the linkage units in NSW and WA were purposely not used during the linkage process, but were used solely to measure linkage quality once the CDL had completed its linkages. By comparing the CDL links with those of the jurisdictions, the CDL was able to evaluate its ability to link very large dataset to a high quality in a short period of time. The results for all linkages were exceptionally high. In total, 99.2% of links found by the CDL were correct, and 96.8% of all links were found. The CDL was successful at closely replicating jurisdictional links in a short time span. The CDL obtained an overall linkage accuracy measure (F-measure) of 0.99 for WA data, and 0.97 for NSW data. Both results were very high. The lower linkage quality obtained for NSW data could be attributed to poorer data quality.

Additional projects utilising cross-jurisdictional linkage infrastructure are in train. These include an exploration of the burden and cost of health care due to injury (which utilises state morbidity, emergency and mortality datasets) and an investigation into the role of perinatal factors in the developmental and educational outcomes of Australian children, (using state level birth and perinatal datasets and the Australian Early Development Index, a national collection on young children's development [23]). The range of possible research projects which can use cross jurisdictional linked data is large and diverse and will have the capacity to improve government policy and planning. The possibility for data linkage research looks set to be restricted only by imagination.

Progress

As results show, the CDL has met its objective of "establishing a secure and efficient data linkage system to facilitate linkage between jurisdictional datasets" [14]. The CDL has established a secure IT environment, instituted strong governance arrangements and implemented a unique cross-jurisdictional operational model. As evidenced by Proof of Concept linkage results, the CDL has also developed the technical capability to undertake large-scale data linkage and produce high-quality linkage output.

Current developments

The CDL is currently continuing with the development of a full production linkage system. In the past, production linkage systems have been limited by their inability to handle increasingly large datasets. The major reason for this poor scalability is the exponential growth in the number of possible matches as so-called "master datasets" extend. To address this and ensure sustainability of national infrastructure, the CDL has designed an efficient and sustainable component-based production linkage system. The system has been designed to securely link event data based on probabilistic matching of demographic information. A new grouping methodology has been implemented that operates at record-pair level. The system has the functionality to support changes in records and datasets over time. Additionally, the linkage system provides functionality to support its own administration by operational staff.

The issues in implementing cross jurisdictional linkage are not only technical. There are also significant challenges around management and governance, engagement with stakeholders, and working in a federated environment with differing legislation. The researchers working with cross jurisdictional linked data also face challenges around merging data from different states and working with different collection methodologies and variable definitions.

Future directions

Data linkage in Australia is an evolving space. At the same time as the PHRN and CDL were developing, a number of Commonwealth government agencies came together to establish a set of guiding principles for data integration involving Commonwealth data [21]. Governance and institutional arrangements for Commonwealth data integration projects have now also been articulated and an accreditation process has recently been put in place.

With safeguards in place, it should be possible to adapt the existing CDL Cross-Jurisdictional Model to accommodate the linkage of State-based datasets to Commonwealth-held data. The resulting infrastructure would provide a resource which can be used to create epidemiological and management information that can be used to investigate and model interactions within a complex, federated Australian health system. Data linkage at this scale would significantly improve Australia's capacity to carry out population health research at a truly national level.

Conclusion

Governments and universities in Australia understand that linked administration data can provide an unparalleled resource for the monitoring and evaluation of services. However, for a number of reasons, these data have not previously been readily available to researchers.

The infrastructure established by the CDL presents a major opportunity to exploit administrative collections and improve the quality of population research data across Australia, with the consequential benefits of improved health and wellbeing of Australians.

Acknowledgements

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network. The authors would like to thank the reviewers for their invaluable comments.

Author details

[1]Curtin University, Perth, Western Australia. [2]CSIRO Mathematics, Informatics and Statistics, Canberra, ACT, Australia. [3]Menzies Research Institute, Tasmania, Australia.

Received: 16 August 2012 Accepted: 21 December 2012
Published: 29 December 2012

References

1. Goldacre M, Glover J (Eds): *The value of linked data for policy development, strategic planning, clinical practice and public health: An international perspective. Symposium on Health Data Linkage*. Adelaide University: Public Health Information Development Unit; 2003.
2. Brook EL, Rosman DL, Holman CDAJ: Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. *Aust N Z J Public Health* 2008, **32**(1):19–23.
3. Hall SE, Holman CDAJ, Finn J, Semmens JB: Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records. *Int J Qual Health Care* 2005, **17**(5):415–420.
4. Sibthorpe B, Kliewer E, Smith L: Record linkage in Australian epidemiological research: health benefits, privacy safeguards and future potential. *Aust J Public Health* 1995, **19**(3):250–256.
5. Hobbs M, McCall M: Health statistics and record linkage in Australia. *J Chronic Disease* 1970, **23**:375–381.
6. Semmens J, Lawrence-Brown M, Fletcher D, Rouse I, Holman CDJ: The Quality of Surgical Care Project: A Model to Evaluate Surgical Outcomes in Western Australia Using Populaiton-Based Record Linkage. *Aust N Z J Surg* 1998, **68**(6):397–403.
7. Roos LL, Wajda A: Record Linkage Strategies: Part 1: Estimating Information and Evaluating Approaches. *Methods Inf Med* 1990, **30**(2):117–123.
8. Gill LE, OX-LINK: *The Oxford Medical Record Linkage System*, Record Linkage Techniques. Oxford: University of Oxford; 1997: p. 19.
9. Kendrick SW, Clarke JA: The Scottish Medical Record Linkage System. *Health Bulletin (Edinburgh)* 1979, **51**:72–79.
10. Holman D, Bass A, Rouse I, Hobbs M: Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Aust N Z J Public Health* 1999, **23**(5):453–459.
11. Ford DV, Jones KH, Verplancke J-P, Lyons RA, John G, Brown G, *et al*: The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009, **9**(1):157. doi: 10.1186/1472-6963-9-157.
12. Holman CDAJ, 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, **32**(4):766–777.
13. Lawrence G, Dinh I, Taylor L: The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation. *Health Inf Manage J* 2008, **37**(2):60–62.
14. NCRIS: *Funding Agreement for the National Collaborative Research Infrastructure Strategy's Research Capability known as 'Population Health Research Network*. Canberra: Commonwealth Department of Education Science and Training; 2009.
15. O'Keefe CM, Ferrante AM, Boyd JH, Semmens JB: *Operational Models 2nd Consultation Draft, Version 0.5*. Perth, WA: Population Health Research Network Centre for Data Linkage; 2009.
16. Kelman CW, Bass AJ, Holman CDJ: Research use of linked health data - a best practice protocol. *Aust N Z J Public Health* 2002, **26**(3):251–255.
17. O'Keefe CM, Ferrante AM, Boyd JH: *National Linkage Keys and National Linkage Map: Ownership and Governance. Draft Version 0.5*. Perth, WA: Population Health Research Network Centre for Data Linkage; 2010.
18. Ferrante A, Boyd JH: *Data Linkage Software Evaluation: A First Report (Part I)*. Perth: PHRN Centre for Data Linkage, Curtin University; 2010.
19. Ferrante AM, Boyd JH: *Data Linkage Software Evaluation: A First Report (Part II) Function and Features*. Perth: PHRN Centre for Data Linkage, Curtin University; 2010.
20. O'Keefe C, Ferrante A, Boyd J: *CDL Operational Model Part 1*. Curtin University: Population Health Research Network Centre for Data Linkage; 2010.
21. Australian Government: *High Level Principles for Data Integration involving Commonwealth Data for Statistical and Research Purposes*. Canberra: Australian Government; 2010.
22. Ferrante A, Boyd J: A transparent and transportable methodology for evaluating Data Linkage software. *J Biomed Inform* 2012, **45**(1):165–172.
23. Goldfeld S, Sayers M, Brinkman S, Silburn S, Oberklaid F: The Process and Policy Challenges of Adapting and Implementing the Early Development Instrument in Australia. *Early Educ Dev* 2009, **20**(6):978–991. cited 2012/11/29.

doi:10.1186/1472-6963-12-480
Cite this article as: Boyd *et al.*: Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC Health Services Research* 2012 **12**:480.

Publication 3

Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., McInneny, K., Brown, A. P., Spilsbury, K., Gillies, M, & Semmens, J. B. (2015). **Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia.** *BMC health services research, 15*(1), 1.

Contribution:

SR supported the development of this paper, carried out the linkage with other co-authors, and made contributions to the final version of the manuscript.

Boyd *et al. BMC Health Services Research* (2015) 15:312
DOI 10.1186/s12913-015-0981-2

BMC Health Services Research

RESEARCH ARTICLE Open Access

Accuracy and completeness of patient pathways – the benefits of national data linkage in Australia

James H. Boyd*, Sean M. Randall, Anna M. Ferrante, Jacqueline K. Bauer, Kevin McInneny, Adrian P. Brown, Katrina Spilsbury, Margo Gillies and James B. Semmens

Abstract

Background: The technical challenges associated with national data linkage, and the extent of cross-border population movements, are explored as part of a pioneering research project. The project involved linking state-based hospital admission records and death registrations across Australia for a national study of hospital related deaths.

Methods: The project linked over 44 million morbidity and mortality records from four Australian states between 1st July 1999 and 31st December 2009 using probabilistic methods. The accuracy of the linkage was measured through a comparison with jurisdictional keys sourced from individual states. The extent of cross-border population movement between these states was also assessed.

Results: Data matching identified almost twelve million individuals across the four Australian states. The percentage of individuals from one state with records found in another ranged from 3-5 %. Using jurisdictional keys to measure linkage quality, results indicate a high matching efficiency (F measure 97 to 99 %), with linkage processing taking only a matter of days.

Conclusions: The results demonstrate the feasibility and accuracy of undertaking cross jurisdictional linkage for national research. The benefits are substantial, particularly in relation to capturing the full complement of records in patient pathways as a result of cross-border population movements.

The project identified a sizeable 'mobile' population with hospital records in more than one state. Research studies that focus on a single jurisdiction will under-enumerate the extent of hospital usage by individuals in the population. It is important that researchers understand and are aware of the impact of this missing hospital activity on their studies.

The project highlights the need for an efficient and accurate data linkage system to support national research across Australia.

Background

Administrative data as a research tool

Administrative datasets are a powerful resource enabling health researchers to answer epidemiological questions that require long-term follow up on large samples of the population [1]. Access to administrative collections such as hospital records, health registries and birth and death information enables research which would otherwise be very expensive and organisationally difficult to undertake [2].

To allow researchers to gain a picture of an individual's health over time, data linkage techniques are utilised to identify which administrative records from multiple datasets belong to the same person. This process allows the researcher to answer questions about the health of individuals over time, rather than solely about discrete health events [3].

Data linkage has several advantages over other study methods. It is far less intrusive and costly than collecting the same information by other means, such as through

* Correspondence: j.boyd@curtin.edu.au
Centre for Population Health Research, Faculty of Health Sciences, Curtin University, Bentley 6102, WA, Australia

large-scale surveys. It allows entire populations to be studied, reducing common problems with follow-up encountered in survey based research designs [4]. Its shortcomings lie in the inflexibility of the data (only information already recorded can be used for analysis). Data linkage studies can also face issues regarding loss to follow up; individuals can move out of a catchment area under study, for instance. The extent of this loss to follow up, and its effect on research results, is largely unknown.

Data linkage methods and linkage quality

In the absence of a unique identifier, data linkage is carried out using demographic information such as name, date of birth and address. As these identifiers can change and be in error (or contain missing information), probabilistic statistical methods are used to ensure the highest quality of linked data [5].

Two types of errors impact linkage quality: false positives, where two records are designated as a match when they should not be, and false negatives, where two records are designated as a non-match when they should not be. The rate of these two errors, measured through precision (or positive predictive value) and recall (sensitivity) statistics, determines overall linkage quality [6].

Ensuring high linkage quality is difficult and typically requires manual efforts. Organisations involved in routine, large-scale data linkage frequently employ a system of manual review of created links to monitor and maintain linkage quality [7, 8]. This can be time and resource intensive, and some errors can still exist even after review. As datasets become larger, the cost and time of manual review becomes prohibitive.

Linkage infrastructure in Australia

Data linkage facilities exist in many parts of the world including Australia, the UK and Canada [4, 9–12]. Australia has been a pioneer in the development of linkage infrastructure for research. Western Australia (WA) has operated a linkage unit since 1995, while the Centre for Health Record Linkage (CHeReL) in New South Wales (NSW) has been in operation since 2006 [13].

From 2009, there has been significant additional government investment in expanding the data linkage research infrastructure in Australia [14]. The creation of a "cross-jurisdictional" linkage capability (that is, the ability to link data from more than one state or territory) was a key component of the Population Health Research Network (PHRN) initiative established under the National Collaborative Research Infrastructure Strategy [15, 16]. Given the federated nature of healthcare service delivery in Australia (that is, some services are delivered and administered at state level, while others are delivered and administered at Commonwealth level), cross-jurisdictional linkage is an essential component of national infrastructure. Without cross-jurisdictional data linkage capabilities, research aimed at national level or targeting issues of common interest (e.g. health service use along border areas) cannot be undertaken. Research at a national level also has other benefits, such as increased statistical power, and reduced loss to follow up caused by interstate movement.

Several 'Proof of Concept' (POC) collaboration projects were initiated by the PHRN to demonstrate the feasibility of moving large datasets across the country, linking these to a high quality in a short period of time, and using the subsequent linked data to answer research questions of national importance [16].

The first of these POC collaborations linked hospital admissions records with death data across several states, focusing on deaths occurring in hospital or within 30 days of hospitalisation. The project was the first of its kind in Australia.

Study aims

The purpose of this paper is twofold. Firstly, to highlight the technical achievements associated with undertaking data linkage for this first POC collaboration.

The paper intends to show that national linkage of 'big data' can be carried out efficiently and accurately. As well as scalable linkage services, an effective national linkage infrastructure needs to deliver high quality linkage results. Current methods for ensuring high linkage quality rely heavily on manual processes, which are not feasible on large datasets. For national linkage to be viable, high linkage quality must be achieved and maintained through automated methods alone.

The second aim of the paper is to demonstrate the importance and impact of cross-jurisdictional linkage. The study will capture population movement at individual or person-based level through linkage of disparate datasets, enabling researchers to assess the full extent of health service utilisation across state borders. The effect of more complete patient pathways on research outcomes has not been previously documented and is not well understood. With reliable estimates of cross-border population flows and service utilisation, researchers can gain a better picture of the need for national linkage studies over state-based linkages projects.

Methods

Datasets and ethics approvals

The data for the POC collaboration included up to ten years of state-based hospital admissions and mortality records from four Australian states between 1st July 1999 and 31st December 2009: Western Australia (WA), New South Wales (NSW), South Australia (SA) and

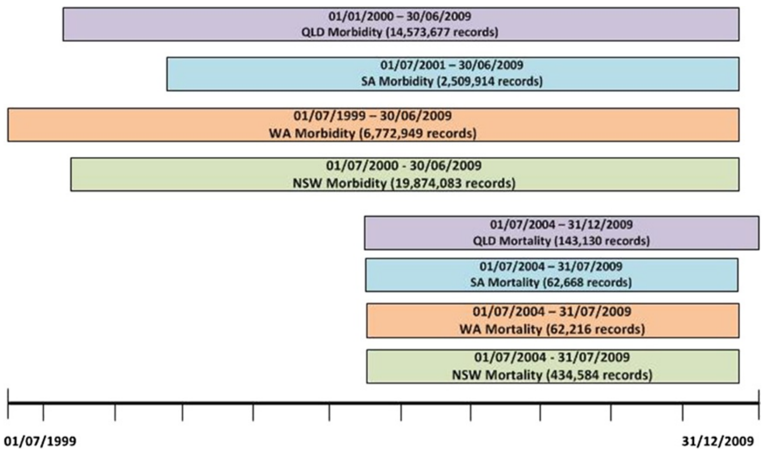


Fig. 1 Datasets provided for proof of concept collaboration

Queensland (QLD) (see Fig. 1). Hospital data was supplied from both public and private hospitals in WA, NSW and QLD; at the time of the project, only admissions from public hospitals in SA were available for linkage. Ethical approval for this study was obtained from Human Research Ethics Committees in WA Health, QLD Health, SA Health, the Cancer Institute NSW and Curtin University (WA).

A total of 44,433,221 records were provided for linkage. In keeping with the separation principle [17], only demographic information was supplied for linkage [16]. Each record comprised information on the person's full name, sex, date of birth and address, as well as admission and separation dates for hospital events (or date of death, for mortality events). Over 30 % of NSW and QLD hospital records did not contain any name information, these records were sourced from private hospitals which did not permit the disclosure of this information. Table 1 provides a summary by state and data collection of the missing data within the variables supplied for linkage.

As WA and NSW had well established linkage infrastructure in place, records from these states had been linked and extensively reviewed *within* their own jurisdiction and assigned a jurisdiction-specific linkage key. These linkage keys identified which records within a particular state belonged to a person within that state. Using these jurisdictional keys, it was possible to directly compare our linkage quality results with those from each of these jurisdictions.

Linkage strategy

Probabilistic linkage methods were used for matching, owing to their flexibility and simplicity [18, 19]. Notwithstanding the size of the datasets, this matching process involved a series of comparisons between two records and a decision as to whether they belong to the same individual. The matching process included a 'blocking' step which limited comparisons to those records which share a minimum level of identifying information. This was important with the large datasets as the potential number of comparisons would be too large to process without the blocking step.

A set of blocking variables were defined for the project [18] and only records which agreed on one of these blocks were compared. The linkage strategy involved two blocks, the first used phonetic surname code (soundex) in combination with first initial and the second

Table 1 Percentage of missing data in linkage variables

Linkage Variables	NSW		WA		SA		QLD	
	Hospital	Mortality	Hospital	Mortality	Hospital	Mortality	Hospital	Mortality
Family name	31.9 %	<0.1 %	<0.1 %	<0.1 %	5.3 %	<0.1 %	34.7 %	<0.1 %
Given name(s)	33.9 %	<1.0 %	<1.0 %	<1.0 %	5.5 %	<0.1 %	36.4 %	<0.1 %
Sex	<0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %
Date of Birth	<0.1 %	<0.1 %	<0.1 %	<1.0 %	<0.1 %	<0.1 %	<0.1 %	<0.1 %
Address	7.5 %	<0.1 %	<1.0 %	2.9 %	8.1 %	<1.0 %	<0.1 %	<0.1 %
Suburb	<1.0 %	1.7 %	<0.1 %	<1.0 %	6.9 %	<1.0 %	<0.1 %	<1.0 %
Postcode	<1.0 %	1.3 %	<1.0 %	<1.0 %	8.5 %	<1.0 %	<0.1 %	4.0 %

selected record pairs for comparison on date of birth and sex [6].

The matching step involved comparing all demographic variables in each blocked pair of records. Each comparison had an associated weight based on the specific agreement and disagreement information provided by individual variables. These variable weights were based on the probability that two values agreed on a record pair given that the two records belong to the same person and the probability of two records belonging to different people when they had the same value.

Agreement and disagreement weights were estimated using knowledge from previous linkages, and refined further in a number of pilot linkages. After computing these weights, a pair comparison score was created by summing agreement and disagreement weights across the demographic variables. If the comparison score for a pair of records exceeded a specified threshold, it was deemed a match [18].

All available demographic variables were used for comparison. Alphabetic variables were compared using the Jaro-Winkler string comparator [20] which computes a score based on the similarity of the strings. Year of birth was scored on a graded scale, receiving a higher score the closer the values were to each other. All other comparisons were based solely on whether the values exactly matched or not.

All datasets were linked to all other datasets, and each dataset was also internally linked. Linkages were initially performed without reference to the provided jurisdictional linkage keys so as to measure linkage quality against these.

Linkage quality

Of primary interest in measuring linkage accuracy is the number of true matches and non-matches identified as links and non-links. To evaluate linkage quality, three standard metrics were used: precision, recall and F-measure [21].

Precision refers to the proportion of returned links that are true matches. It is sometimes referred to as positive predictive value. Recall is the proportion of all true matches that have been correctly linked. Recall is also known as sensitivity. The F-measure of a linkage is the harmonic mean between precision and recall. This provides a single figure with which linkage quality can be compared.

These metrics have been highlighted as suitable for measuring data linkage quality [22, 23] and have been used in evaluations of linkage software [6].

Following the assessment of linkage accuracy, a series of automated and semi-automated procedures were used on the patient based record groups to identify and resolve errors. These included algorithms which addressed groups with multiple deaths, hospital records after death as well as unusually large groups (i.e. groups with more than 5000 records).

Linkage efficiency

As a cross jurisdictional project, which involved data files with large number of records, it was not feasible to compare all possible record pairs to establish links. Instead a series of blocks were employed which aimed to reduce the number of comparisons without having an impact on linkage quality (i.e. reduce comparisons without missing ‘True Positive’ links). To assess the efficiency and quality of the blocks we calculated two complexity metrics, the reduction ratio and pairs completeness score [24].

The reduction ratio provided an assessment of the decrease in comparisons as a result of the blocking strategy. This was calculated as the ratio of actual blocked comparisons to the total possible comparisons and measured the efficiency of the strategy without measuring the impact on linkage quality.

The percentage of ‘true pairs’ blocked or pairs completeness metric measured the number of true positive pairs compared in the blocking strategy as a percentage of all possible true positive pairs identified using the jurisdictional linkage keys for WA and NSW records. Records from these states were used as they have been linked and extensively reviewed *within* their own jurisdiction.

There is an obvious balance between the reduction ratio and percentage of ‘true pairs’ blocked. If the comparisons are reduced for efficiency it can have an impact on linkage quality and increasing comparisons to maximise quality can significantly impact the time required to process the linkage. The blocking strategy is therefore the reference point for all additional linkage quality estimates (i.e. precision and recall).

Results

Over 44 million records across morbidity and mortality collections were linked within and between each jurisdiction. The linkage strategy produced a series of records pairs each with a matching score which were used to identify records belonging to an individual across all data sources. The linkage strategy was evaluated in terms of blocking efficiency and linkage quality.

Blocking efficiency

Using the blocking strategy outlined, approximately 142 billion comparisons were performed during the linkage process. These matching assessments made up only 0.014 % of all possible record pairs from the full comparison space. The blocking process was similar within each jurisdiction, with the state-based reduction ratio

Table 2 Blocking efficiency

Linkage Comparison Summary	NSW	WA	SA	QLD	Total
Number of records supplied for linkage:					
Hospital	19,874,083	6,772,949	2,509,914	14,573,677	43,730,623
Mortality	434,584	62,216	62,668	143,130	702,598
Total	20,308,667	6,835,165	2,572,582	14,716,807	44,433,221
Linkage comparison space:					
Blocked Comparisons	26,071,726,251	6,328,711,086	821,279,963	13,597,405,294	142,112,536,420
Reduction Ratio	0.99987	0.99973	0.99975	0.99987	0.99986
Possible Pairs Blocked (%)	0.0126 %	0.0271 %	0.0248 %	0.0126 %	0.0144 %
'True' Pairs Blocked (%)[a]	99.76 %	99.95 %	-	-	-

[a]'True' pairs based on the jurisdictional linkage key supplied by WA and NSW

ranging between 0.99973 and 0.99987. Table 2 provides a summary of the matching comparisons undertaken.

Linkage accuracy

Linkage results were compared against those produced by state-based linkage units in WA and NSW (both these datasets were supplied with a jurisdictional linkage key). The jurisdictional links from these states were used as a gold standard and allowed an evaluation of linkage quality against each individual state (that is, comparing within-state results only).

The accuracy results for all linkages were exceptionally high with over 99.76 % of all 'true pairs' made available for comparison through blocking i.e. a very small number of pairs identified by WA and NSW jurisdictional linkage keys were lost as a result of the blocking strategy (Table 2). This provided a baseline for assessing the linkage quality of all blocked comparisons.

In WA, over 99.9 % of the morbidity pairs identified as links were found to be correct, and 98.1 % of all possible within-jurisdiction morbidity links were found. This resulted in a maximum F-measure quality score of 0.99 where 1.000 would indicate a perfect linkage (see Table 3) indicating 'an average' error rate for morbidity data from these jurisdictions of less than 1 %.

One factor which had an effect on both blocking and matching accuracy was missing data in the linkage variables (Table 1). Over 30 % of NSW hospital records did not contain any name information (these records were sourced from private hospitals which did not permit release of this information). As a consequence, the quality results for our linkages on WA data were higher than that of NSW. The linkage of morbidity records in NSW provided an overall F-measure of 0.976 (precision = 98.8 % and recall = 96.3 %).

NSW results were further disaggregated by hospital status (public versus private). Records from public hospitals showed much higher results (F-Measure = 0.995) indicating that the lack of demographic information accounted for the drop in linkage quality (Table 3).

Table 3 Linkage quality

Jurisdictional Data	NSW			WA
	Morbidity	Public	Private	Morbidity
Accuracy of national linkage:				
Precision	0.988	0.994	0.983	0.999
Recall	0.963	0.996	0.917	0.981
F-measure[a]	0.976	0.995	0.949	0.990

[a]F-measure is the harmonic mean of precision and recall

Patient summary statistics

The final results of the linkage across the various jurisdictions are summarised in Table 4. Across the four jurisdictions almost 12 million individuals accounted for the 44 million records. Under half (45 %) of the individuals identified with hospital records had a single hospital admissions record; with the remainder having an average of 5.9 hospital records per person.

The number of individuals with a single hospital record varied across the four jurisdictions with Western Australia (WA) having the smallest proportion (35 %) and South Australia (SA) having the highest (52 %). Similarly, the average group size (i.e. the record per individual) varied between 6.2 and 5.2 in WA and SA respectively. It should be noted that the South Australian figures do not include private hospital records which may influence the proportion of singleton groups in that state.

Cross-border population movements and hospital usage statistics over the study period are summarised in Table 5. The proportions of individuals in each state with records in one or more of the other three states were classified as a 'mobile' population. The 'mobile' population was largest in QLD with 5 % of individuals having hospital records in other states and lowest in SA

Table 4 Patient summary results

Linkage Results - Summary	NSW	WA	SA	QLD	Total
Number of individuals:					
Identified from Hospital and Death records	5,796,784	1,558,999	848,446	3,995,812	11,954,874
Hospital events within individual groups:					
Number of individuals hospitalised	5,782,670	1,554,313	833,781	3,979,562	11,907,114
Singleton hospital records[a]	2,598,149	544,484	433,277	1,831,768	5,407,678
%	44.9 %	35.0 %	52.0 %	46.0 %	45.4 %
Maximum number of hospital records	2,297	2,245	2,393	2,393	2,393
Average group size[b]	5.4	6.2	5.2	5.9	5.9

[a] Individuals who only have one hospital record in their group
[b] Singletons are not included in the total number of individuals for this calculation

and WA where 3 % were classified as 'mobile' individuals. The 'mobile' population accounted for between 4 and 7 % of the episodes of care in each state jurisdiction.

Discussion

The linkage described here was part of a large POC collaboration that tested the efficiency and accuracy of newly established national data linkage infrastructure in Australia.

Linkage quality

The accuracy and efficiency of the linkage was shown to be high with a large number of 'blocked' pairs comparisons removed from the matching process with very little impact on the linkage quality. Using validated linkage information from WA and NSW, little discrepancy was found between the created links and those found by jurisdictional linkage units in those states. The existence of some discrepancies can be attributed to the additional quality work carried out by those jurisdictional linkage units. Jurisdictional linkage units in Australia typically employ extensive manual review of created links, along with stringent regular manual quality checks. Further errors are identified through feedback following the use of the linked data in research projects. Some of the difference in results could also be attributed to the limited number of identifiers supplied for cross-jurisdictional linkage. Linkage quality depends heavily upon the quality of the underlying dataset. NSW data, with one third of names missing, had the lowest overall linkage quality using our linkage strategy (without additional data collections or clerical intervention).

These quality comparisons rely on the use of jurisdictional linkages as the gold standard. These links from WA and NSW have been validated by researchers who have used them widely. In addition, significant expertise has been developed by these organisations which have a long history of linkage. Having access to two entire sets of extensively checked links allowed us to gain a very accurate estimate of our quality. Few previous investigations into linkage quality have had such a reliable and large gold standard with which to test their results. Typical measures of linkage quality have used samples of links to gain an estimate of quality, often able only to estimate the number of incorrect links created, with the number

Table 5 Patient mobility

	NSW	WA	SA	QLD
Population mobility or cross-border flows (over study period)				
Mobile population[a]	205,551	47,575	29,645	202,859
% of individuals in that state	4 %	3 %	3 %	5 %
Static population[b]	5,591,233	1,511,424	818,801	3,792,953
% of individuals in that state	96 %	97 %	97 %	95 %
Number of events				
Mobile population	1,135,905	248,480	137,234	1,014,912
% of jurisdiction records	6 %	4 %	5 %	7 %
Static population	19,172,762	6,586,685	2,435,348	13,701,895
% of jurisdiction records	94 %	96 %	95 %	93 %

[a] Mobile population refers to the number of individuals in a jurisdiction/state that have records in other states
[b] Static population refers to the number of individuals in a jurisdiction/state that have records *only* in that state

of links missed essentially unknown [25], or have used relative measures to estimate missed links [26] which allows relative comparison, but not absolute quality measures.

Cross border population movement

Linking hospital records across four states over a ten year time span showed that, on average, between 3 % and 5 % of patients within one state had hospital record in another state. The results further showed that between 4 % and 7 % of hospital records occurring in a state can be attributed to an individual who also has records in another state.

These findings suggest that research studies examining patient pathways may underestimate the total number of event records belonging to individuals if they do not factor in cross-border hospital admissions. In studies involving hospital admissions events from a single state, it is important that researchers are aware of the incomplete nature of information and the impact this may have on research outcomes. The size and impact of this underestimation will depend on several factors such as the selection of study cohort and the study period, with longer study periods being more susceptible to population movement into and out of the jurisdiction.

It has been shown that data linkage quality can have an overall impact on research outcomes, potentially biasing results [27]. However, incomplete patient pathways as a result of cross-border flows are not often addressed in linked epidemiological research. When a significant proportion of patients are having hospital activity in more than one jurisdiction, it is important that researchers understand the impact of this incomplete information on single jurisdiction studies [28]. The impact of this data omission on research outcomes is uncertain and warrants further research into the effect of linkage quality and incomplete patient pathways on research outcomes.

Conclusion

These results show the feasibility of large scale data linkage infrastructure, producing high quality results through efficient linkage processes. Overall, data linkage quality in large scale linkage remains very high, despite the lack of stringent manual quality review procedures, which would be extremely costly on datasets of this size. Importantly, this type of linkage identifies cross-border population movement, enabling researchers to fully describe patient pathways.

The national linkage infrastructure has been successfully used to join together records from multiple administrative datasets which belong to the same person. The infrastructure has been developed to be flexible and scalable, addressing the traditional challenges and limitations of efficiently linking national data. With an increasingly 'mobile' population with life event records in different states, this "cross-jurisdictional" linkage service will have positive benefits on Australian health research.

Competing interests

The authors declare they have no competing interests.

Authors' contributions

Linkage design provided by JHB, AMF and JBS. Initial linkage model developed and refined by SMR and KM. Further technical design provided by JKB, APB and KS. Linkage quality assessment and interventions carried out by SMR, JKB, MG, AMF and JHB. First draft of manuscript provided by JHB; subsequently edited by SMR, AMF and KS. All authors have approved the final version of the manuscript.

Acknowledgements

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy and Super Science Initiative's Population Health Research Network. The project would not have been possible without the support of the data linkage units who helped coordinate access to the jurisdictional data.

Received: 24 December 2014 Accepted: 29 July 2015
Published online: 08 August 2015

References

1. Virnig BA, McBean M. Administrative data for public health surveillance and planning. Annu Rev Public Health. 2001;22(1):213–30.
2. Sibthorpe B, Kliewer E, Smith L. Record linkage in Australian epidemiological research: health benefits, privacy safeguards and future potential. Aust J Public Health. 1995;19(3):250–6.
3. Holman D, Bass A, Rouse I, Hobbs M. Population-based linkage of health records in Western Australia: Development of a health services research linked database. Aust N Z J Public Health. 1999;23.
4. Holman CDAJ, 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;32(4):766–77.
5. Newcombe H, Kennedy J. Record linkage: making maximum use of the discriminating power of identifying information. Commun ACM. 1962;5(11):563–6.
6. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. J Biomed Inform. 2012;45(1):165–72.
7. Quality Assurance [http://www.cherel.org.au/quality-assurance]
8. Rosman D, Garfield C, Fuller S, Stoney A, Owen T, Gawthorne G: Measuring data and link quality in a dynamic multi-set linkage system. In: Symposium on Health Data Linkage (https://www.adelaide.edu.au/phidu/publications/pdf/1999-2004/symposium-proceedings-2003/rosman_a.pdf): 20–21 March 2002 2002; Sydney; 2002: 4.
9. Kendrick SW, Clarke JA. The Scottish Medical Record Linkage System. Health Bulletin (Edinburgh). 1979;51:72–9.
10. Gill LE. OX-LINK: The Oxford Medical Record Linkage System. In: Record Linkage Techniques. Oxford: University of Oxford; 1997. p. 19.
11. Roos LL, Wajda A. Record Linkage Strategies: Part 1: Estimating Information and Evaluating Approaches. Winnipeg: University of Manitoba; 1990. p. 28.
12. Field K, Kosmider S, Johns J, Farrugia H, Hastie I, Croxford M, et al. Linking data from hospital and cancer registry databases: should this be standard practice? Internal medicine journal. 2010;40(8):566–73.
13. Lawrence G, Dinh I, Taylor L. The Centre for Health Record Linkage: A New Resource for Health Services Research and Evaluation. Health Information Management Journal. 2008;37(2):60–2.
14. NCRIS. Funding Agreement for the National Collaborative Research Infrastructure Strategy's Research Capability known as 'Population Health Research Network'. Canberra: Commonwealth Department of Education Science and Training; 2009.
15. Frommer PM, Madronio C, Kemp S, Jenkin R, Reitano R. NCRIS Capability 5.7: Population Health and Data Linkage. Sydney: University of Sydney; 2007. p. 8.

16. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC Health Serv Res. 2012;12.
17. Kelman C, Bass A, Holman D. Research use of linked health data: A best practice protocol. Aust N Z J Public Health. 2002;26:5.
18. Newcombe HB. Handbook for Record Linkage: Methods for Health and Statistical Studies, Administration and Business. New York: Oxford University Press; 1988.
19. Jaro MA. Probabilistic Linkage of Large Public Health Data Files. Stat Med. 1995;14:491–8.
20. Jaro MA. "UNIMATCH: A record linkage system: User's manual", Technical Report, US Bureau of the Census, Washington D.C. 1976.
21. Christen P, Goiser K. Assessing Deduplication and Data Linkage, Quality: What to Measure. In: Proceedings of the Fourth Australasian Data Mining Conference Sydney; 2005: 16.
22. Christen P, Goiser K. Quality and Complexity Measures for Data Linkage and Deduplication. In. Canberra: Department of Computer Science, Australian National University; 2004.
23. Bishop G, Khoo J. Methodology of Evaluating the Quality of Probabilistic Linking. Canberra: Australian Bureau of Statistics, Analytical Services Branch; 2007. p. 20.
24. Christen P, Goiser K. Quality and complexity measures for data linkage and deduplication. Quality Measures in Data Mining. Berlin Heidelberg: Springer; 2007. 127–151.
25. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. 2010.
26. Campbell KM, Deck D, Krupski A. Record linkage software in the public domain: a comparison of Link Plus, The Link King and a 'basic' deterministic algorithm. Health Informatics. 2008;14(1):5–15.
27. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Med Res Methodol. 2014;14(1):36.
28. Harron K, Wade A, Muller-Pebody B, Goldstein H, Gilbert R. Opening the black box of record linkage. J Epidemiol Community Health. 2012;66(12):1198–8.

Chapter 2

Uses of record linkage

Research Output

Supporting Publications

4. **Randall, S. M.**, Zilkens, R., Duke, J. M., & Boyd, J. H. (2016). **Western Australia population trends in the incidence of acute myocardial infarction between 1993 and 2012**. *International Journal of Cardiology, 222*, 678-682.

5. Duke, J. M., Rea, S., Boyd, J. H., **Randall, S. M.**, & Wood, F. M. (2015). **Mortality after burn injury in children: a 33-year population-based study**. *Pediatrics, 135(*4), e903-e910.

6. Duke, J. M., Boyd, J. H., Rea, S., **Randall, S. M.**, & Wood, F. M. (2015). **Long-term mortality among older adults with burn injury: a population-based study in Australia.** *Bulletin of the World Health Organization, 93*(6), 400-406.

7. **Randall, S. M.**, Fear, M. W., Wood, F. M., Rea, S., Boyd, J. H., & Duke, J. M. (2015**). Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study**. *BMJ open*, *5*(9), e009395.

2.1. Applications of record linkage

The value of record linkage comes not from the creation of a linkage map itself (a listing of which records belong to which individuals), but from the use this linkage map is put to in practice. This can take numerous forms. In business contexts, record linkage is often used to remove duplicates from customer-based lists [69], while governments often link administrative information for use in reporting and planning [70]. Along with these uses, record linkage is an important research tool in the area of epidemiology. By linking together different administrative data collections, such as hospital admission and emergency presentation records, disease registries along with birth and death records, a detailed picture emerges of an individual's lifelong health. As administrative records typically capture an entire population, this data allows researchers to answer numerous health questions. These include questions on the nature of diseases (disease prevalence and incidence estimates, long term morbidity and survival following particular diseases, risk factors for specific diseases), treatments (impact of introduced treatments on disease incidence, morbidity and survival), health policy evaluation (the impact of introduced health policy) and health service utilisation (investigating the effective use of health resources).

Analysis of linked administrative data is less widely used in other research fields, although this is beginning to change. Both educational datasets (used to measure the impact of policy changes on educational outcomes [71]) and criminal justice datasets (linking together police, court and prison records for recidivism research [52, 72]) have seen recent use. Efforts to bring all these separate sectors together for research are currently underway [73].

The key benefit to the use of administrative linked data is in the ability to answer questions about entire populations without the prohibitive cost and time commitment of conducting large scale surveys. The use of record linkage techniques for health and other research is mainly limited by the nature of the collected information. The primary purpose of these datasets is administrative, and as such they do not always contain the in-depth clinical or service information desired by researchers.

The possibilities of record linkage for health research are highlighted further in the below examples.

2.2. The use of record linkage for health research: two examples

Understanding the overall trends of common health conditions (for instance to see whether their occurrence is increasing or decreasing over time) is a routine public health requirement, vital for determining the importance of public health interventions. While disease trends can often be carried out using hospital morbidity collections, linked data can provide a more accurate clinical picture. For instance, one study utilised hospital morbidity separations data to investigate how the incidence of acute myocardial infarctions in Australia was changing over time. They found the incidence of acute myocardial infarction in Australia to be increasing [74]; this stood in contrast to results from other Western nations, which uniformly showed a decline in incidence rates [75].

The use of linked hospital and mortality data provides a different picture of the incidence of acute myocardial infarction. Firstly, linked data (as opposed to counts of individual separations) allows us to take into account ward and hospital transfers by patients – these result in additional separations, and so will result in a single myocardial infarction being 'double counted' if they were transferred. Individuals who are discharged home may also come back into hospital with complications of their myocardial infarction; in this case, they may again receive the same primary diagnosis, and so are at risk of being double counted. International guidelines suggest ignoring any readmissions that occur within 30 days of a first myocardial infarction, as these are likely not a separate occurrence [76]. Again, knowing which separation belongs to which person, linked data can achieve this. Finally, by using hospital morbidity data linked with mortality data we can include in the study those who suffered from an acute myocardial infarction, but died before they could be admitted to hospital.

By counting acute myocardial infarctions in this way, a different picture emerges; the overall acute myocardial infarction rate is decreasing, in line with other Western nations. The use of linked data has provided more accurate information, which in this case has resulted in a change in overall conclusions. This study is described further in *Publication 4: Western Australia population trends in the incidence of acute myocardial infarction between 1993 and 2012.*

As well as using linked data to monitor and understand trends in disease, linked data can be used to generate new knowledge and hypotheses about health conditions. An example of this is an investigation into burn injury. This research program was initiated due to the absence of long-term follow-up data regarding people who experienced burn injuries, specifically whether they were more likely to suffer particular health conditions in the years after burn injury. This was hypothesised in light of evidence of long term persistence of systemic inflammatory responses after both minor and severe burn injury [77].

Linked data provided an excellent opportunity to explore this issue. Using linked data, all individuals who were hospitalised for a burn injury in Western Australia over 32 years (1980 to 2012) could be identified and followed up over this time period. For each individual with a burn injury, hospital admissions occurring after the burn injury (along with mortality records) could be used to gain a picture of that individual's health after the injury. To determine whether those with a burn had higher morbidity, they needed to be compared to a control group, who did not experience a burn. Using linked data, this is relatively straightforward. The Western Australian data linkage system includes, along with hospital morbidity, the electoral roll (voting in compulsory in Australia, so this is an almost complete listing of all adults in the state), and birth registry information dating back to the 1970s; as such, it should cover close to the whole of the Western Australian population. In this study, for each individual with a burn injury, four controls were chosen of the same gender who were born in the same five year band, and who were still alive at the time of burn injury.

This study aimed to compare individuals with a burn to those without, to see the likelihood of subsequent morbidity. Other individual factors may effect this likelihood; for instance, history of smoking, or previous medical history; this information may not be equally distributed between the burn and control cohort. Factors such as these can be controlled for using statistical techniques within the analysis. However, they can only be controlled for if we are aware of them and they are recorded accurately. Smoking status for instance, is not recorded within any hospital data collection, and so we cannot control for this factor. On the other hand, previous medical history can be obtained by looking at medical records for an individual from before their burn injury.

This research program has resulted in several publications It found significantly higher mortality rates for those with a burn compared to controls, for both children (*Publication 5: Mortality after burn injury in children: a 33-year population-based study* [78]) and adults (*Publication 6: Long-term mortality among older adults with burn injury: a population-based study in Australia [79])*, after controlling for demographic characteristics and health status. In addition, this research found people with burns have higher rates of admission for cardiovascular diseases [80, 81], as well as musculoskeletal conditions (*Publication 7: Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study [82]*). It has been hypothesised that these changes were caused by a systemic response to burn injury, in particular persistent elevated levels of catecholamines. Further research based on these results has shown cardiovascular changes in a mouse model of burn injury, along with visible heart differences in burn patients found on echocardiography [6]. These results confirm that pathophysiological changes due to burn injury are the likely cause of the increased cardiovascular hospitalisations found when analysing linked data.

These two examples serve to illustrate the scope of the studies utilising linked data; in reality, they only scratch the surface of potential research questions. However, for these and other important questions to be answered, robust linkage methods are required, which can ensure quality and reduce privacy risk.

Publication 4

Randall, S. M., Zilkens, R., Duke, J. M., & Boyd, J. H. (2016). **Western Australia population trends in the incidence of acute myocardial infarction between 1993 and 2012.** *International Journal of Cardiology, 222*, 678-682.

Contribution:

SR developed the methodology and research design, reviewed the literature, performed all analyses, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

Publication 5

Duke, J. M., Rea, S., Boyd, J. H., **Randall, S. M.**, & Wood, F. M. (2015). **Mortality after burn injury in children: a 33-year population-based study**. *Pediatrics, 135(*4), e903-e910.

Contribution:

SR supported the development of this paper, conducting the analysis with other co-authors, and making contributions to the final version of the manuscript.

Publication 6

Duke, J. M., Boyd, J. H., Rea, S., **Randall, S. M.**, & Wood, F. M. (2015). **Long-term mortality among older adults with burn injury: a population-based study in Australia.** *Bulletin of the World Health Organization, 93*(6), 400-406.

Contribution:

SR supported the development of this paper, conducting the analysis with other co-authors, and making contributions to the final version of the manuscript.

Publication 7

Randall, S. M., Fear, M. W., Wood, F. M., Rea, S., Boyd, J. H., & Duke, J. M. (2015**). Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study**. *BMJ open*, *5*(9), e009395.

Contribution:

SR supported the development of this paper, conducting the analysis, interpreting results, writing the result section of the manuscript, and making contributions to the final version of the manuscript.

BMJ Open

Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study

Sean M Randall,[1] Mark W Fear,[2] Fiona M Wood,[2,3] Suzanne Rea,[2,3]
James H Boyd,[1] Janine M Duke[2]

To cite: Randall SM, Fear MW, Wood FM, *et al*. Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study. *BMJ Open* 2015;**5**:e009395. doi:10.1136/bmjopen-2015-009395

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2015-009395).

Received 17 July 2015
Revised 31 July 2015
Accepted 18 August 2015

[1]Centre for Data Linkage, Curtin University, Perth, Western Australia, Australia
[2]Burn Injury Research Unit, School of Surgery, University of Western Australia, Perth, Western Australia, Australia
[3]Burns Service of Western Australia, Royal Perth Hospital and Princess Margaret Hospital, Perth, Western Australia, Australia

Correspondence to
Professor Janine M Duke;
janine.duke@uwa.edu.au

ABSTRACT

Objective: To investigate if adults who are hospitalised for a burn injury have increased long-term hospital use for musculoskeletal diseases.

Design: A population-based retrospective cohort study using linked administrative health data from the Western Australian Data Linkage System.

Subjects: Records of 17 753 persons aged at least 20 years when hospitalised for a first burn injury in Western Australia during the period 1980–2012, and 70 758 persons who were age and gender-frequency matched with no injury admissions randomly selected from Western Australia's electoral roll.

Main outcome measures: Admission rates and cumulative length of stay for musculoskeletal diseases. Negative binomial and Cox proportional hazards regression modelling were used to generate incidence rate ratios (IRR) and HRs with 95% CIs, respectively.

Results: After adjustment for pre-existing health status and demographic characteristics, the burn cohort had almost twice the hospitalisation rate for a musculoskeletal condition (IRR, 95% CI 1.98, 1.86 to 2.10), and spent 3.70 times as long in hospital with a musculoskeletal diagnosis (95% CI 3.10 to 4.42) over the 33-year period, than the uninjured comparison cohort. Adjusted survival analyses of incident post-burn musculoskeletal disease admissions found significant increases for the 15-year post burn discharge period (0–6 months: HR, 95% CI 2.51, 2.04 to 3.11; 6 months–2 years: HR, 95% CI 1.77, 1.53 to 2.05; 2–15 years: HR, 95% CI 1.32, 1.23 to 1.42). Incident admission rates were significantly elevated for 20 years post-burn for minor and severe burn injury for a range of musculoskeletal diseases that included arthropathies, dorsopathies, osteopathies and soft tissue disorders.

Conclusions: Minor and severe burn injuries were associated with significantly increased post-burn incident admission rates, long-term hospital use and prolonged length of stay for a range of musculoskeletal diseases. Further research is required that facilitates identification of at-risk patients and appropriate treatment pathways, to reduce the long-term morbidity associated with burns.

Strengths and limitations of this study

- Population-based linked administrative health data provide a cost-effective means to examine long-term health impacts of burn injury.
- Population-based linked administrative health data minimise issues of selection and reporting bias, and loss to follow-up.
- The retrospective longitudinal study design included a comparison group.
- Lack of individual-based risk factor data.
- Small proportion of patient population with severe burns limited examination of long-term effects.

INTRODUCTION

Despite advances in surgical and medical treatment, burn injuries continue to present significant challenges to clinicians, and to burn survivors, often leading to long-term psychological and physical impairments.[1] It is well documented that severe burn injury induces acute inflammatory and hypermetabolic responses that persist for at least 2 years after the initial injury.[2 3] The subsequent metabolic demands and energy requirements are profound and induce mobilisation of proteins and amino acids, resulting in an associated increase in protein turnover, degradation and negative nitrogen balance, characteristics of serious illness.[4] The physiological impacts of hypermetabolism include protein catabolism, losses of body weight, lean body mass and bone mineral density.[5 6]

In addition to the hypermetabolic response and muscle wasting there is extensive and sustained suppression of bone formation as a result of the systemic inflammatory and endocrine responses to severe burn injury.[7–9] Muscle wasting and immobility after burn injury can also directly alter the balance of bone synthesis and degradation, leading to bone loss.[7 9–11] Burn patients are also at risk of vitamin D deficiency, contributing to bone loss.[12–15] Vitamin D deficiency may develop progressively as a result of a number of factors including prolonged sun avoidance during treatments,[16] hypoparathyroidism[17 18] and low serum levels of cholesterol evident after burn injury,[19 20] preventing the synthesis and activation of vitamin D. In adults and children, loss of bone mass and changes in bone density predispose burn patients to an increased incidence of fractures and potentially to lifelong issues associated with osteoporosis.[7 11 14 15 21 22]

Minor burn injuries represent the majority of burn related hospital admissions in developed countries,[23 24] and there is a growing interest in understanding the potential for systemic responses after non-severe burns. Recent research has demonstrated long-term impacts after minor burn injury on bone marrow,[25] muscle, innervation and bone,[26–28] with population-based findings of increased cancer incidence[29 30] and long-term mortality.[31 32] Non-severe burn injury has also been found to have a sustained impact on reducing trabecular bone density long after resolution of inflammation.[27] However, the clinical relevance of these bone alterations is not yet clear. Investigation of the long-term effects of severe and minor burn injury is important to provide evidence for improvements in acute burn care.

To date, limited data have been available to examine the long-term health impacts of burns. Population-based linked health administrative data provide a cost-effective means to examine long-term morbidity trends expressed in the number of hospital admissions and length of stay for specific disease classifications.[33] Given the profound growth and musculoskeletal changes experienced during childhood and adolescence with the potential for different post-burn impacts, this study was limited to data of adult burn survivors 20 years and above. The aim of this study was to use population-based linked health administrative data to determine if adults hospitalised for a burn injury have increased long-term hospital use for musculoskeletal diseases, after adjustment for sociodemographic factors and pre-existing comorbidities.

METHODS

Our study formed part of the Western Australian Population-based Burn Injury Project—a retrospective cohort investigation—that uses administrative health data from the Western Australian Data Linkage System (WADLS), a validated linkage system that links several core datasets for the entire population of Western Australia.[34] The project was approved by the human research ethics committees of the University of Western Australia and the Western Australian Department of Health.

Analyses were performed on a de-identified extraction of hospital morbidity records for all individuals who were aged at least 20 years when admitted to a hospital in Western Australia with a first burn injury between 1 January 1980 and 30 June 2012, undertaken by the WADLS. A first (index) burn injury was defined as the first hospital admission in a patient's medical record in which a burn injury was given as the principal diagnosis or an additional diagnosis, defined by International Classification of Diseases and Related Health (ICD) 9 CM 940–949 or ICD10 AM T20–T31. A population-based comparison cohort was randomly selected from Western Australia's electoral roll. Any person with an injury hospitalisation during the study period was excluded from this cohort by WADLS staff. The resultant comparison cohort was frequency matched (4:1) on birth year and sex of each burn injury case for each year from 1980 to 2012. Cohort selection and analytical methods have been reported previously.[31]

Data from Western Australia's Hospital Morbidity Data System and Death Register were linked to the burn and non-injured cohorts for the period 1980–2012. Hospital admissions data included principal and additional diagnoses, age at admission, sex, and Aboriginal status, date of admission, date of discharge or other separation and mode of separation. Data supplied for the burn and non-injured cohorts also included geocoded place of residence and geocoded indices of geographical remoteness[35] and social disadvantage.[36] Geographical remoteness was classified into five categories: major cities, inner regional, outer regional, remote and very remote. The social disadvantage index was reclassified into quintiles (most to least disadvantaged). The mortality data included date of death and cause of death.

An individual listed as Aboriginal or Torres Strait Islander on any admission record was categorised as Aboriginal. Supplementary codes ICD9-CM 948 or ICD10-AM T31 were used to classify the patients into those with minor burns (<20% of total body surface area (TBSA)), severe burns (≥20% TBSA) and burns of unspecified TBSA. Comorbidity (baseline) was assessed, with a 5-year look-back period, using the Charlson Comorbidity Index (CCI) and the principal and additional diagnoses included in the hospital morbidity records (0 CCI=0; 1 CCI≥1).[37 38] A record of an existing congenital anomaly was identified using principal and additional diagnosis data (ICD9 740–759; ICD10 Q00-G99). The final discharge date for the index burn admission was used as the study start for follow-up for the burn cases and the respective frequency matched non-injury controls.

Categorical and non-parametric continuous variables were compared using χ^2 and Kruskal Wallis tests, respectively. A p value of 0.05 or lower was considered statistically significant. The total number of admissions for musculoskeletal disorders after burn injury discharge and the cumulative length of stay for principal diagnosis musculoskeletal disorders classified by subchapter headings were used as outcome measures (ICD10 AM: M00-M25—arthropathies (including infectious and inflammatory arthropathies, arthrosis and other joint disorders); M30-M36—systemic connective tissue disorders; M40-M54—dorsopathies (including deforming spinal conditions, spondylopathies and other disorders of the spine); M60-M79—soft tissue disorders (including disorders of muscles, synovium and tendons, other soft tissue disorders); M80-M94—osteopathies and chrondropathies (including disorders of bone density, osteopathies and chrondropathies). ICD-10 codes were mapped to ICD-9 codes.[39] The hospitalisation of the first burn injury was not included in these outcomes. Crude yearly admission rates were calculated for these variables. Adjusted rate ratios (incidence rate ratio (IRR) and 95% CI) between the burn injury and no injury cohorts and the outcome measures were generated using negative binomial regression. Sociodemographic (gender, Aboriginality, 5-year age group, social disadvantage, remoteness of residence and year of admission) and health status information (comorbidity at baseline, history of musculoskeletal disease, congenital anomaly) were included as covariates in the models to adjust for potential confounding.

Survival analyses of incident hospital use for arthropathies, systemic connective tissue disorders, dorsopathies, soft tissue disorders, osteopathies and chrondropathies (ie, ICD subchapter headings) as well as specific subsets of musculoskeletal diseases, infectious arthropathy, inflammatory polyarthropathy and disorders of bone density and structure, were conducted using multivariate Cox proportional hazards models. Analyses were conducted on the burn and uninjured cohorts excluding those with prior admission for musculoskeletal disease, and additionally, excluding those with a record of an injury admission in the burn cohort. Cox models were adjusted for sociodemographic and health status variables (as above). Analyses were undertaken of subgroups by burn TBSA severity and gender-specific analyses were undertaken to explore potential gender dimorphic differences in disease incidence. The proportional hazard assumption for the burn injured versus non-injured was tested using scaled and unscaled Schoenfeld residuals and by adding a group-by-time interaction term.[40] Where preliminary analyses showed non-proportionality, adjusted HR and 95% CI for first record of musculoskeletal admission for burn versus non-injury cohorts were modelled for time periods guided by Aalen's linear hazard models and plots.[41]

Attributable risk percentages (AR%) were calculated as the adjusted rate ratio (IRR, HR) minus one, divided by the adjusted rate (IRR, HR) ratio, multiplied by 100.[42] AR% was used to estimate the proportion of long-term and incident hospital use for musculoskeletal diseases, where burn injury was a component cause.[43] Statistical analyses were performed using Stata V.12 (StataCorp. LP, College Station, Texas, USA).

RESULTS

The burn injury cohort included 17 753 persons aged at least 20 years when hospitalised for a first burn injury during the period January 1980–June 2012. The median age was 36 (IQR: 26–51) and 71.1% were male. Four per cent had severe burns of 20% TBSA or greater, 49% had burns of less than 20% TBSA and for 47%, the TBSA was unspecified. Eighteen per cent of the burn cohort had sustained full thickness burns, 37% partial thickness, 19% erythema and 29% had experienced burns for which the burn depth was unspecified; an individual may have had multiple burns sites and depths recorded. Among the burn cohort, 34% had a record of a non-burn injury admission (before or after index burn), and 13.8% had a previous musculoskeletal hospital admission. The burn injury cohort had a median follow-up time of 13.6 years (IQR; minimum (min)–maximum (max): 5.5–22.7; >0–32.5) for a total of 250 670 person years.

The comparison non-injury cohort comprised 70 758 persons with median age 41 (IQR: 27–51) and males accounting for 71.0%. The uninjured cohort had a median follow-up time of 14.5 years (IQR; min–max: 6.9–23.1; 0.01 to 32.5) for a total of 1 067 568 person years. Refer to table 1 for other baseline sociodemographic and health status variables for the burn and non-injury cohorts. The burn injury cohort comprised significantly higher proportions of Aboriginal people, people who were socially disadvantaged, people living in regions outside of major cities and people who had pre-existing comorbidity when compared with the non-injured comparison cohort.

Admissions for diseases of musculoskeletal system—rates and cumulative length of stay

There were 10 761 hospital admissions occurring after burn hospitalisation discharge with a primary diagnosis of a musculoskeletal disease. Arthropathies were the most common cause of musculoskeletal admission, followed by dorsopathies (see table 2). A total of 55 810 days were spent in hospital for musculoskeletal diagnosis after a burn hospitalisation. The median length of hospital stay was 1 day (IQR: 0–5 days). The total number of days spent in hospital with a musculoskeletal primary diagnosis by the uninjured cohort was 68 946, where the median length of stay of musculoskeletal admissions was 1 day (IQR: 0–3 days). The length of stay of 0 days represents admission and discharge occurring on the same day.

Unadjusted incidence rates for musculoskeletal admissions and lengths of stay are shown in figure 1. These graphs show a higher rate of admissions for musculoskeletal diseases, and more time spent in hospital with musculoskeletal diseases for burn patients compared with uninjured patients over the entire 33-year period.

After adjustment for pre-existing health and sociodemographic characteristics, the burn cohort had almost twice as many hospitalisations for a musculoskeletal condition (IRR, 95% CI 1.98, 1.86 to 2.10), and spent 3.70 times as long in hospital with a musculoskeletal diagnosis (95% CI 3.10 to 4.42) over the 33-year period, than

Table 1 Baseline demographic and pre-existing health status factors for the burn injury (total) and the frequency matched non-injury (total) cohorts

Characteristics	No injury N (%)	Burn injury N (%)	p Value
Total	70 758	17 753	
Demographic			
Aboriginality			
Yes	809 (1.1)	2139 (12.0)	<0.001
Social disadvantage quintiles*			
Quintile 1 (most disadvantaged)	8307 (11.8)	3716 (21.4)	<0.001
Quintile 2	15 623 (22.1)	5536 (31.8)	
Quintile 3	12 930 (18.3)	3681 (21.2)	
Quintile 4	12 988 (18.4)	2207 (12.7)	
Quintile 5 (least disadvantaged)	20 738 (29.4)	2255 (13.0)	
Remoteness†			
Major city	53 358 (75.6)	8965 (51.5)	<0.001
Inner regional	6564 (9.3)	1954 (11.2)	
Outer regional	6036 (8.6)	2783 (16.0)	
Remote	2789 (4.0)	1913 (11.0)	
Very remote	1839 (2.6)	1803 (10.4)	
Health status			
Comorbidity†			
Any (Charlson Comorbidity Index, CCI≥1)	4,132 (5.8)	2833 (16.0)	<0.001
Prior admission for disease of musculoskeletal system§	3889 (5.5)	2455 (13.8)	<0.001

*Socio-economic Indexes for Areas (SEIFA) socioeconomic disadvantage quintiles; missing values 2% burn, 0.2% no injury.
†Accessibility Remoteness Index for Australia (ARIA+) remoteness classification; missing values 1.9% burn, 0.02% no injury.
‡Based on CCI using 5-year look-back.
§Principal diagnosis record of hospitalisation for musculoskeletal diseases (ICD9 710–739; ICD10 M00-M99) using 5-year look-back period.
ICD, International Classification of Diseases and Related Health.

Table 2 Number of admissions (%) for musculoskeletal diseases classified by ICD subchapter codes in the non-injury and burn injury cohorts

Musculoskeletal subconditions	Number of admissions (%)	
	No injury	**Burn injury**
Total	**20 223**	**10 761**
Arthropathies	**9513 (47.0)**	**4362 (40.5)**
Infectious arthropathy	82 (0.4)	153 (1.4)
Inflammatory polyarthropathy	620 (3.1)	795 (7.4)
Osteoarthritis	3706 (18.3)	1346 (12.5)
Other joint disorders	5105 (25.2)	2068 (19.2)
Dorsopathies	**5620 (27.8)**	**3584 (33.3)**
Spondylopathies	1571 (7.8)	747 (6.9)
Other dorsopathies	3918 (19.4)	2747 (25.5)
Osteopathies and chrondropathies	**772 (3.8)**	**764 (7.1)**
Disorders of bone density and structure	366 (1.8)	355 (3.3)
Soft tissue disorders	**3809 (18.8)**	**1747 (16.2)**
Disorder of muscles	95 (0.5)	79 (0.7)
Disorders of synovium/ tendons	833 (4.1)	324 (3.0)
Other soft tissue disorders	2881 (14.2)	1344 (12.5)
Connective tissue disorders	**143 (0.7)**	**64 (0.6)**
Other	**366 (1.8)**	**243 (2.3)**

ICD, International Classification of Diseases and Related Health.

the uninjured comparison cohort. The adjusted AR% suggested that the burn cohort experienced an excess of 49.5% of admissions for musculoskeletal conditions (n=5329) and 73.0% of all days spent in hospital for musculoskeletal conditions (n=40 727 days), when compared to the uninjured cohort.

Unadjusted admission rates for the musculoskeletal system sub-categories are shown in figure 2. After controlling for demographic factors and previous health status, those who had a burn had higher rates of admissions for arthropathies (IRR, 95% CI 1.64, 1.54 to 1.75), dorsopathies (IRR, 95% CI 2.16, 1.93 to 2.41), osteopathies and chondropathies (IRR, 95% CI 5.64, 4.56 to 6.97) and soft tissue disorders (IRR, 95% CI 1.91, 1.76 to 2.06). There was no difference in the rate of admissions for connective tissue disorders between the burn and the non-injured cohorts (IRR, 95% CI 0.88, 0.48 to 1.62). The adjusted AR% suggested that the burn cohort experienced excess post-burn admissions of 39.0% for arthropathies (n=1702), 53.7% for dorsopathies (n=1925), 82.3% for osteopathies and chrondropathies (n=629), and 47.6% for soft tissue disorders (n=832), when compared with the uninjured cohort.

Examination of more specific subconditions showed a large and significant increase in the rate of admissions for disorders of bone density (IRR, 95% CI 13.87, 9.89 to 19.44), along with increases in infectious arthropathies (IRR, 95% CI 1.98, 1.46 to 2.69) and inflammatory

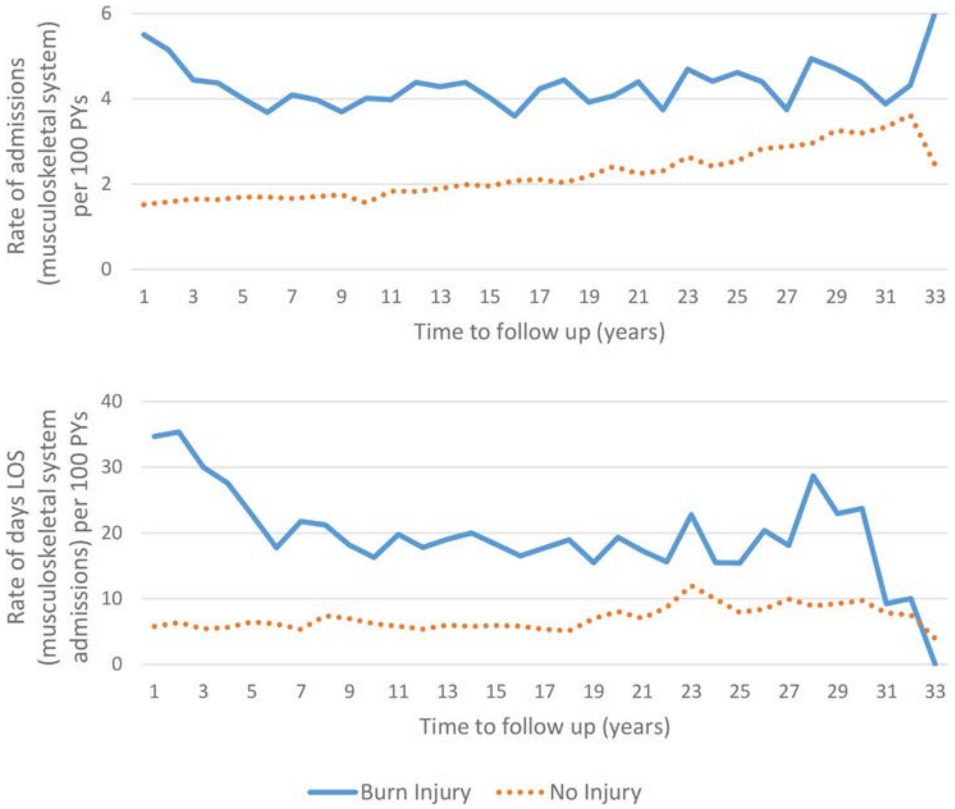


Figure 1 Unadjusted rates of hospital admissions and cumulative length of stay (per 100 person-years (PYs)) for musculoskeletal diseases (total) among adults with burn injury versus no injury.

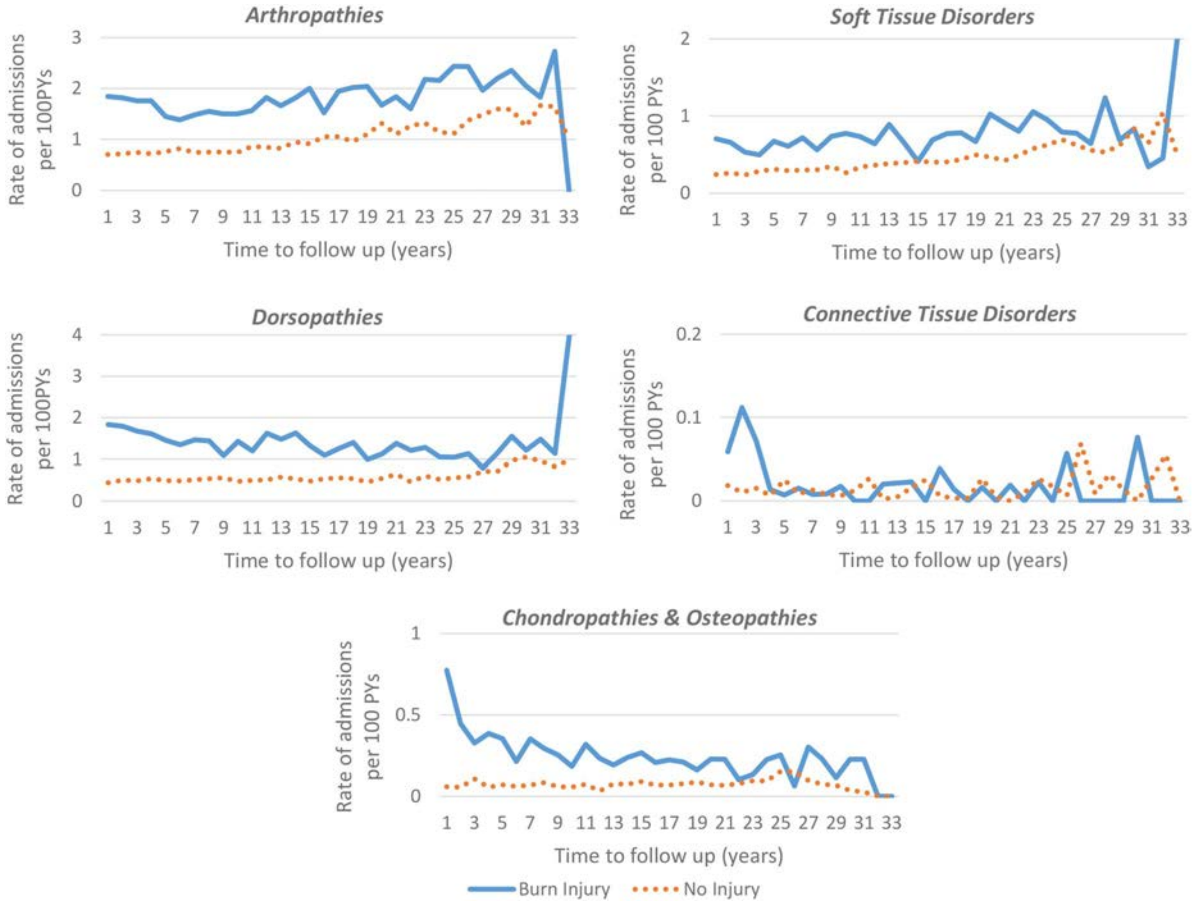


Figure 2 Unadjusted rates (per 100 person-years (PYs)) of hospital admissions by musculoskeletal disease subgroup for adults with burn injury versus no injury.

polyarthropathies (IRR, 95% CI 3.82, 2.61 to 5.59) for those hospitalised with a burn injury.

A similar picture was shown for the length of time spent in hospital, with longer cumulative length of stay after a burn hospitalisation with arthropathies (IRR, 95% CI 1.64, 1.54 to 1.75), dorsopathies (IRR, 95% CI 2.16, 1.93 to 2.41), osteopathies and chondropathies (IRR, 95% CI: 5.64, 4.56 to 6.97), and soft tissue disorders (IRR, 95% CI 1.91, 1.76 to 2.06). Those with a burn injury spent almost 50 times longer in hospital with disorders of bone density compared to the uninjured cohort (IRR, 95% CI 49.24, 28.72 to 84.42), with increases in cumulative time spent in hospital also found for infectious arthropathies (IRR, 95% CI 13.1, 5.62 to 30.56) and inflammatory polyarthropathies (IRR, 95% CI 1.98, 1.46 to 2.69). No significant difference was found between the burn and uninjured cohorts for length of time in hospital with connective tissue disorders (IRR, 95% CI 0.65, 0.23 to 1.85).

Analysis by TBSA severity (see table 3) showed that increased hospitalisation rates were associated with severe and minor burns, with those with minor burns having higher admission rates for arthropathies, dorsopathies, osteopathies and chondropathies, and soft tissue disorders when compared with the uninjured cohort.

Incidence—survival analysis

Analysis of time until first musculoskeletal admission (post-burn) was performed on the burn and uninjured cohorts who did not have a prior record of musculoskeletal hospitalisation, with the additional exclusion from the burn cohort of those with a record of a principal diagnosis injury admission. There were 10 440 individuals in this reduced burn cohort, of which 1779 had a first time (incident) musculoskeletal admission within the study period. The uninjured comparison cohort comprised of 66 869 controls, of which 9689 had an incident musculoskeletal admission in this study period.

Results of multivariate Cox regression modelling for an incident musculoskeletal (combined diseases) admission revealed evidence of non-proportionality. First time musculoskeletal (combined diseases) admissions were most frequent immediately after burn and while remaining significantly elevated, decreased over the study period. Adjusted analyses conducted on partitioned time windows found significant HRs for the first 6 months after burn (HR, 95% CI 2.51, 2.04 to 3.11), from 6 months to 2 years after burn (HR, 95% CI 1.77, 1.53 to 2.05), and from 2 to 15 years after burn (HR, 95% CI 1.32, 1.23 to 1.42). Differences were found for severe (HR, 95% CI 4.37, 2.32 to 8.25), minor (HR, 95% CI 2.25, 1.68 to 3.01) and unknown TBSA (HR, 95% CI 2.42, 1.74 to 3.37) for the first 6 months; only minor burns (HR, 95% CI 6 months–2 years; 1.68, 1.39 to 2.04, 2–15 years; 1.44, 1.30 to 1.59) and unknown TBSA burns (HR, 95% CI 6 months–2 years; 1.83, 1.54 to 2.17, 2–15 years; 1.28, 1.17 to 1.41) showed increased admissions over the first 15 years.

After adjustment for confounders, both males (HR, 95% CI 0–6 months; 2.62, 2.03 to 3.39, 6 months–2 years; 1.83, 1.54 to 2.17, 2–15 years; 1.31, 1.21 to 1.42) and females (HR, 95% CI 0–6 months; 2.09, 1.35 to 3.23, 6 months–2 years; 1.68, 1.27 to 2.23, 2–15 years; 1.34, 1.17 to 1.53) with burns showed significant increases over the first 15 years post-burn, compared with uninjured males and females, respectively.

In the burn cohort, there were 968 first time admissions for arthropathies, 468 for dorsopathies, 155 for osteopathies and chrondropathies (n=70 bone density disorders), 541 for soft tissue disorders and 8 first time admissions for connective tissue disorders. First time admissions for infectious arthropathy and inflammatory polyarthropathy accounted for 24 and 67 of the total incident arthrropathy admissions, with median (IQR) times to incident admission of 8.4 years (3.4–16.3) and 9.4 years (3.8–16.9), respectively. Results for adjusted Cox regression models for incident musculoskeletal subconditions are shown in table 4. Evidence of non-proportionality was common for numerous subconditions, with higher rates of admissions in the burn cohort

Table 3 Adjusted IRR and 95% CIs for recurrent admissions musculoskeletal subconditions, by burn severity compared with the uninjured cohort

Musculoskeletal subconditions	Severe burns* IRR (95% CI)†	Minor burns* IRR (95% CI)	Unspecified severity* IRR (95% CI)
Arthropathies	**1.54 (1.20 to 1.99)**	**1.56 (1.41 to 1.73)**	**1.71 (1.58 to 1.85)**
Infectious arthropathy	6.20 (1.54 to 24.87)	8.05 (4.37 to 14.82)	2.23 (1.49 to 3.35)
Inflammatory polyarthropathy	2.91 (1.22 to 6.91)	1.99 (1.21 to 3.28)	2.06 (1.48 to 2.86)
Dorsopathies	**1.69 (1.06 to 2.68)**	**1.98 (1.68 to 2.33)**	**2.38 (2.07 to 2.73)**
Osteopathies and chondropathies	**3.32 (1.80 to 6.10)**	**8.44 (6.27 to 11.36)**	**3.97 (3.18 to 4.95)**
Disorders of bone density and structure	5.09 (0.82 to 31.50)†	22.22 (13.08 to 37.74)	9.89 (6.89 to 14.19)
Soft tissue disorders	**2.41 (1.70 to 3.42)**	**1.99 (1.77 to 2.24)**	**1.86 (1.69 to 2.05)**
Connective tissue disorders	**(no admissions)**	**0.98 (0.38 to 2.54)**	**0.92 (0.44 to 1.90)**

*Severe: TBSA 20%+; minor TBSA<20%; unspecified: TBSA unknown.
†All models were adjusted for sociodemographic (age group, gender, Aboriginal status, social disadvantage, remoteness), index year and health (comorbidity, prior musculoskeletal admission) status.
IRR, incidence rate ratios; TBSA, total body surface area.

Table 4 Adjusted HRs for first time post-burn admissions for musculoskeletal subconditions, comparing the burn cohort with the uninjured cohort

Musculoskeletal subconditions†	HR (95% CI)*	Attributable risk %*	Number of admissions attributable to burn injury
Arthropathies			
0–1 year after burn	2.02 (1.58 to 2.57)	50.5	47
1–20 years after burn	1.26 (1.17 to 1.37)	20.6	154
Infectious arthropathy			
0–20 years after burn	2.34 (1.33 to 4.10)	57.3	11
Inflammatory polyarthropathy			
0–33 years after burn	1.68 (1.27 to 2.21)	40.5	27
Dorsopathies			
0–20 years after burn	1.39 (1.24 to 1.55)	28.1	114
Osteopathies and chondropathies			
0–1 year after burn	6.99 (3.94 to 12.41)	85.7	23
1–5 years after burn	3.08 (2.12 to 4.48)	67.5	31
5–20 years after burn	1.69 (1.29 to 2.23)	40.8	28
Disorders of bone density and structure			
0–2 years after burn	7.14 (4.11 to 12.40)	86.0	11
2–33 years after burn	1.78 (1.27 to 2.51)	43.8	18
Soft tissue disorders			
0–10 years after burn	1.74 (1.52 to 1.99)	42.5	120
10–33 years after burn	1.24 (1.08 to 1.42)	19.4	50
Total			567

*Models used data for time after burn discharge (study start) and all models were adjusted for sociodemographic (age group at index, gender, Aboriginal status, social disadvantage, remoteness), index year and comorbidity.
†No analysis for connective tissue disorders due small numbers.

more immediately after the burn injury. All subconditions showed significantly higher rates of first admissions in the first 20 years after burn. The AR% calculated suggested that 567 first admissions to hospital for musculoskeletal diseases could be attributed to burn injury. The largest proportion of these incident admissions attributable to the burn injury were for arthropathies (35.4%), followed by diseases of soft tissue (30.0%), dorsopathies (20.0%) and finally, osteopathies and chrondropathies (14.5%).

DISCUSSION

This study quantifies the increased population-based long-term hospital use for musculoskeletal disorders experienced by adults hospitalised with burn injury, after controlling for demographic and pre-existing comorbidities. Members of the burn injury cohort had 1.98 times the rate of hospitalisations and 3.70 times the length of stay in hospital for combined musculoskeletal disorders post-burn when compared with the uninjured cohort. Arthropathies and dorospathies combined accounted for 74% of all readmissions (prevalent and incident disease admissions) for musculoskeletal disorders. After controlling for sociodemographic and pre-existing health status, the burn cohort experienced significantly higher rates of hospitalisations post-burn for arthropathies, dorsopathies, osteopathies and chrondropathies and soft tissue disorders. No difference was found for admission rates for diseases of the connective tissue when compared with the uninjured cohort.

Assessment of the impact of burn severity on recurrent admissions for musculoskeletal diseases identified significantly elevated rates of similar magnitude for severe and minor burn injury for arthropathies, dorsopathies and soft tissue disorders. However, while admission rates for osteopathies and chrondropathies were significantly elevated for severe burns (3.32 times higher), minor burns were associated with an admisson rate of 8.44 times that observed for the uninjured cohort over the study period. This lack of a dose–response relationship may in part be due to the small numbers of burn survivors with severe burns in this patient population, the TBSA classification used and/ or a 'healthy' survivor effect. An additional possibility is that patients with severe burns have supportive treatments to mitigate systemic inflammatory and endocrine responses and are more likely to have higher levels of continued post-burn care that may lead to earlier diagnoses and management of secondary pathologies, resulting in reduced hospitalisations for more serious presentations. The impact of minor burn injury on musculoskeletal disease admissions is interesting in light of recent evidence of less severe burn causing depletion of trabecular bone volume.[27]

To exclude any potential additive systemic impacts associated with other non-burn injury among members of the burn cohort,[43 44] first time admissions for musculoskeletal diseases post-burn were examined excluding those in the burn cohort with a principal diagnosis injury admission. After adjustment for sociodemographic and health factors, significantly elevated rates of first time post-burn hospitalisations for musculoskeletal disorders were identified. The adjusted AR% for incident post-burn musculoskeletal disease admissions suggested that 567 first time admissions experienced in the burn cohort after discharge, could be attributed to burn injury. Gender-specific analyses found males and females with burn injury to have significantly elevated admission rates.

Long-term admissions for infectious and inflammatory arthropathies were assessed as a means to investigate the potential for chronic impacts of the initial systemic inflammatory and immune response to burn injury.[45–47] Significantly increased incident admission rates for infectious and inflammatory arthropathies post-burn were identified with effects persisting for at least 20 years, suggesting that burn patients are at increased long-term susceptibility to infectious and inflammatory diseases. The increases in inflammatory and infectious diseases may be related to the immune suppression induced by burn injury;[46 48 49] however, recent evidence also suggests long-term changes in the immune response post-burn, including reduced dendritic cell activation and inability of dendritic cells to activate T cells.[44] While the findings of considerably elevated incident admission rates for bone density disorders during the first 2 years post-burn were consistent with other research,[7 11 14 15 21 22] the burn cohort continued to experience significantly elevated incident admissions over the 33-year study period. Burn injury affects multiple systems of the body including the immune, metabolic and endocrine systems, and the pathophysiological mechanisms that underlie the different musculoskeletal conditions discussed in this paper are most likely variable. Future research will be important to identify how burn injury triggers different mechanisms in order to inform early intervention and prevent these long-term pathologies.

Strengths and limitations

Through the use of linked administrative health data, we performed a large scale population-based study with a long follow-up time, accurate pre- and post burn injury measures,[50] a non-injured comparison group and sample size sufficient for quantitative analyses. The analytical strategy of the study was based on the assumption that after controlling for confounders, any excess in hospital use for musculoskeletal diseases in the burn cohort (compared to the non-injured), was an outcome principally associated with the incident burn. We were able to include variables of health, social disadvantage and geographic access to services, to determine the individual burden of hospital readmissions for musculoskeletal disorders and quantify at the population level. While health administrative datasets do not include variables of risk-taking behaviour, analyses undertaken of the burn cohort excluding members with a record of a non-burn injury admission and potentially those of high risk-taking behaviour, also found significantly elevated incident musculoskeletal admission rates. Likewise, models were adjusted for social disadvantage, a factor previously correlated with lifestyle risk factors (eg, nutrition, smoking, alcohol)[51 52] that could be associated with the burn exposure and the outcomes measured. Changes in ICD coding and incomplete TBSA% data may have limited a complete understanding of burn severity on long-term hospital use; however, significantly increased admission rates were observed for severe and minor burns and burns of unspecified TBSA. Any differential effect of burn injury on the incidence of fractures could not be evaluated in this study due to the exclusion of any injury admissions (including fractures) in the comparator uninjured cohort. Further research is planned that will include a non-burn injury cohort where differences in long-term musculoskeletal morbidity will be examined, including fractures.

The out-of-state migration for Western Australia is consistently low at levels below 3%; such losses were not anticipated to bias the results.[53] This study represents burn injury and musculoskeletal diseases serious enough to require hospitalisation and the results may underestimate the impact of burn injury experienced in the community. Research of functional outcomes and quality-of-life of burn patients in Western Australia have provided valuable data on patients during the first year after discharge;[54 55] however, this study has generated new information on the longer-term musculoskeletal morbidity experienced long after healing of the burn wound and cessation of attendance at outpatient burn clinics. Further work that links individuals' pharmaceutical and or primary care data with hospital data will provide a clearer picture of the time of diagnosis of secondary musculoskeletal pathologies and treatment pathways post-burn. We expect that these results would be generalisable to other countries of similar demographics and healthcare systems.

After adjustment of confounders, increased rates of first time admissions and readmissions for musculoskeletal diseases were identified for those with severe and minor burns when compared with a non-injured cohort. These findings highlight long-term effects of burn injury, including minor burn injury, on musculoskeletal morbidity. Further research is required that facilitates identification of at-risk patients, the mechanisms that may be responsible for these long-term effects and appropriate treatment pathways, to reduce the long-term morbidity associated with burns.

Acknowledgements The authors thank the staff of the Health Information Linkage Branch for access to the Western Australian Data Linkage System and for their assistance in obtaining the data, the WA Health Data Custodians for access to the core health datasets and the Western Australian Department of Health. The project was supported by the Raine Medical Research Foundation (Priming grant JMD). JMD is supported by Woodside Corporation. MWF is supported by Chevron Australia.

Contributors The manuscript, including related data, figures and tables has not been previously published and that the manuscript is not under consideration elsewhere. All authors have made contributions to the paper and authorised the submission: JMD designed the study, supported data analyses and interpretation and drafted manuscript. SMR provided data management, analyses and data interpretation. JHB provided statistical advice. MWF, SR and FMW contributed clinical interpretation. All authors contributed to manuscript preparation and critical revision.

Funding Project data costs were supported by a Raine Medical Research Foundation Priming grant (JMD). JMD is supported by Woodside corporate sponsorship and the Fiona Wood Foundation. MWF is supported by Chevron Australia and the Fiona Wood Foundation.

Competing interests None declared.

Ethics approval This study has received ethics approval from the University of Western Australia and the Western Australian Department of Health. Being a large population cohort study using de-identified linked data, approval included a waiver of informed consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Jeschke MG, Chinkes DL, Finnerty CC, *et al*. Pathophysiologic response to severe burn injury. *Ann Surg* 2008;248:387–401.
2. Przkora R, Barrow RE, Jeschke MG, *et al*. Body composition changes with time in pediatric burn patients. *J Trauma* 2006;60:968–71.
3. Przkora R, Jeschke MG, Barrow RE, *et al*. Metabolic and hormonal changes of severely burned children receiving long-term oxandrolone treatment. *Ann Surg* 2005;242:384–9, discussion 390-1.
4. Chang DW, DeSanti L, Demling RH. Anticatabolic and anabolic strategies in critical illness: a review of current treatment modalities. *Shock* 1998;103:155–60.
5. Rennie MJ. Muscle wasting in the muscular dystrophies. *Dev Med Child Neurol* 1985;27:524–7.
6. Rennie MJ. Muscle protein turnover and the wasting due to injury and disease. *Br Med Bull* 1985;41:257–64.
7. Klein GL. Burn-induced bone loss: importance, mechanisms, and management. *J Burns Wounds* 2006;5:e5.
8. Klein GL, Wolf SE, Goodman WG, *et al*. The management of acute bone loss in severe catabolism due to burn injury. *Horm Res* 1997;48(Suppl 5):83–7.
9. Leblebici B, Sezgin N, Ulusan SN, *et al*. Bone loss during the acute stage following burn injury. *J Burn Care Res* 2008;29:763–7.
10. Klein GL, Bi LX, Sherrard DJ, *et al*. Evidence supporting a role of glucocorticoids in short-term bone loss in burned children. *Osteoporos Int* 2004;15:468–74.
11. Klein GL, Herndon DN, Goodman WG, *et al*. Histomorphometric and biochemical characterization of bone following acute severe burns in children. *Bone* 1995;17:455–60.
12. Garrel D. Burn scars: a new cause of vitamin D deficiency? *Lancet* 2004;363:259–60.
13. Klein GL. The interaction between burn injury and vitamin D metabolism and consequences for the patient. *Curr Clin Pharmacol* 2008;3:204–10.
14. Klein GL. Burns: where has all the calcium (and vitamin D) gone? *Adv Nutrition* 2011;2:457–62.
15. Klein GL. Does vitamin D deficiency contribute to post-burn bone loss? *F1000Res* 2012;1:57.
16. Norman AW. Sunlight, season, skin pigmentation, vitamin D, and 25-hydroxyvitamin D: integral components of the vitamin D endocrine system. *Am J Clin Nutr* 1998;67:1108–10.
17. Klein GL, Langman CB, Herndon DN. Persistent hypoparathyroidism following magnesium repletion in burn-injured children. *Pediatr Nephrol* 2000;14:301–4.
18. Murphey ED, Chattopadhyay N, Bai M, *et al*. Up-regulation of the parathyroid calcium-sensing receptor after burn injury in sheep: a potential contributory factor to postburn hypocalcemia. *Crit Care Med* 2000;28:3885–90.
19. Gottschlich MM, Alexander JW. Fat kinetics and recommended dietary intake in burns. *JPEN J Parenter Enteral Nutr* 1987;11:80–5.
20. Gottschlich MM, Jenkins M, Warden GD, *et al*. Differential effects of three enteral dietary regimens on selected outcome variables in burn patients. *JPEN J Parenter Enteral Nutr* 1990;14:225–36.
21. Mayes T, Gottschlich M, Scanlon J, *et al*. Four-year review of burns as an etiologic factor in the development of long bone fractures in pediatric patients. *J Burn Care Rehabil* 2003;24:279–84.
22. Edelman LS, McNaught T, Chan GM, *et al*. Sustained bone mineral density changes after burn injury. *J Surg Res* 2003;114:172–8.
23. Akerlund E, Huss FR, Sjoberg F. Burns in Sweden: an analysis of 24,538 cases during the period 1987–2004. *Burns* 2007;33:31–6.
24. Duke J, Wood F, Semmens J, *et al*. A 26-year population-based study of burn injury hospital admissions in Western Australia. *J Burn Care Res* 2011;32:379–86.
25. Rea S, Giles NL, Webb S, *et al*. Bone marrow-derived cells in the healing burn wound—more than just inflammation. *Burns* 2009;35:356–64.
26. Anderson JR, Zorbas JS, Phillips JK, *et al*. Systemic decreases in cutaneous innervation after burn injury. *J Invest Dermatol* 2010;130:1948–51.
27. O'Halloran E, Kular J, Xu J, *et al*. Non-severe burn injury leads to depletion of bone volume that can be ameliorated by inhibiting TNF-alpha. *Burns* 2015;41:558–64.
28. Morellini NM, Fear MW, Rea S, *et al*. Burn injury has a systemic effect on reinnervation of skin and restoration of nociceptive function. *Wound Repair Regen* 2012;20:367–77.
29. Duke J, Rea S, Semmens J, *et al*. Burn Injury and cancer risk: a state-wide longitudinal study. *Burns* 2011;38:340–7.
30. Duke JM, Bauer J, Fear MW, *et al*. Burn injury, gender and cancer risk: population-based cohort study using data from Scotland and Western Australia. *BMJ Open* 2014;4:e003845.
31. Duke JM, Boyd JH, Rea S, *et al*. Long-term mortality among older adults with burn injury: a population-based study in Australia. *Bull World Health Organ* 2015;93:400–6.
32. Duke JM, Rea S, Boyd JH, *et al*. Mortality after burn injury in children: a 33-year population-based study. *Pediatrics* 2015;135:e903–10.
33. Hendrie D, Miller TR. Assessing the burden of injuries: competing measures. *Inj Control Saf Promot* 2004;11:193–9.
34. Holman CD, Bass AJ, Rouse IL, *et al*. Population-based linkage of health records in Western Australia: development of a health service research linked database. *Aust NZ J Public Health* 1999;23:453–9.
35. Glover J, Tennant S. *Remote areas statistical geography in Australia: notes on the Accessibility/Remoteness Index for Australia (ARIA+ version)*. Working Papers Series No. 9. Adelaide: Public Health Information Development Unit, Adelaide, The University of Adelaide, 2003.
36. Trewin D. *Socio-economic indexes for areas (information paper, census of population and housing)*. Canberra: Australian Bureau of Statistics, 2003.
37. Charlson ME, Pompei P, Ales KL, *et al*. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373–83.
38. Preen DB, Holman CDAJ, Spilsbury K, *et al*. Length of comorbidity lookback period affected regression model performance of administrative health data. *J Clin Epidemiol* 2006;59:940–6.
39. Ministry of Health—Manatu Haurora NZ. Mapping between ICD-10 and ICD-9. Secondary Mapping between ICD-10 and ICD-9. 30 May 2015. http://www.health.govt.nz/nz-health-statistics/data-references/mapping-tools/mapping-between-icd-10-and-icd-9
40. Hosmer DW, Lemeshow S. *Applied survival analysis: regression modeling of time to event data*. New York: Wiley, 1999.
41. Hosmer DW, Royston P. Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model. *Stata J* 2002;2:331–50.
42. Gordis L. *Epidemiology*. 2nd edn. Philadelphia: W.B. Saunders Company, 2000.
43. Cameron CM, Purdie DM, Kliewer EV, *et al*. Ten-year health service use outcomes in a population-based cohort of 21,000 injured adults: the Manitoba injury outcome study. *Bull World Health Organ* 2006;84:802–10.
44. Valvis SM, Waithman J, Wood FM, *et al*. The Immune Response to Skin Trauma is Dependent on the Etilology of Injury in a Mouse Model of Burn and Excision. *J Invest Dermatol* 2015;135:2119–28.
45. Mace JE, Park MS, Mora AG, *et al*. Differential expression of the immunoinflammatory response in trauma patients: burn vs. non-burn. *Burns* 2012;38:599–606.
46. O'Sullivan ST, O'Connor TP. Immunosuppression following thermal injury: the pathogenesis of immunodysfunction. *Br J Plast Surg* 1997;50:615–23.

47. Rock KL, Lai JJ, Kono H. Innate and adaptive immune responses to cell death. *Immunol Rev* 2011;243:191–205.
48. D'Arpa N, Accardo-Palumbo A, Amato G, *et al*. Circulating dendritic cells following burn. *Burns* 2009;35:513–18.
49. Patenaude J, D'Elia M, Hamelin C, *et al*. Selective effect of burn injury on splenic CD11c(+) dendritic cells and CD8alpha(+)CD4(-) CD11c(+) dendritic cell subsets. *Cell Mol Life Sci* 2010;67:1315–29.
50. Department of Health Western Australia. Clinical Information Audit Program—Hospital Activity Report. Operational Directive OD 0201/09. Perth Department of Health WA, 2009.
51. Mishra G, Ball K, Patterson A, *et al*. Socio-demographic inequalities in the diets of mid-aged Australian women. *Eur J Clin Nutr* 2005;59:185–95.
52. Siahpush M, Borland R, Scollo M. Prevalence and socio-economic correlates of smoking among lone mothers in Australia. *Aust NZ J Public Health* 2002;26:132–5.
53. Clark A, Preen DB, Ng JQ, *et al*. Is Western Australia representative of other Australian States and Territories in terms of key socio-demographic and health economic indicators? *Aust Health Rev* 2010;34:210–15.
54. Grisbrook TL, Elliott CM, Edgar DW, *et al*. Burn-injured adults with long term functional impairments demonstrate the same response to resistance training as uninjured controls. *Burns* 2013;39:680–6.
55. Grisbrook TL, Reid SL, Edgar DW, *et al*. Exercise training to improve health related quality of life in long term survivors of major burn injury: a matched controlled study. *Burns* 2012;38:1165–73.

Chapter 3

A review of the literature on record linkage methodology

3.1. Recent developments in record linkage methodology: a survey of the literature

As discussed in Chapter 1, the task of record linkage is complex. Methods which ensure high linkage quality are sought to ensure validity of research results, and reduce the considerable burden that manual methods entail. Timely approaches are required due to the large processing overhead. Issues of privacy remain a concern to many members of the public.

This chapter aims to provide a critical review of the recent literature in record linkage methods. To gain a general sense of the overall trends and focus of the literature, papers on record linkage methodology published in the previous 21 years (between 1995 and 2015) were sourced from Google Scholar and classified based on their content[1].

Figure 4 plots the number of publications on record linkage methodology over time. There has been a large increase in record linkage methodology research over the last 21 years, with the number of papers increasing more than three-fold. There appear two periods of increase; a large increase occurring around 2002, followed by a smaller increase around 2009. These correspond to identifiable trends within the literature. The year of 2002 saw the beginning of a burgeoning interest in machine learning methods for record linkage (some 13 papers were published on this topic from 2002-2004, compared to only one in the previous seven years). At the same time, the field of entity resolution was emerging, with numerous publications devoted to this development. The increase in publications from 2009 appears to result from an increased focus on methods for privacy

[1] The search terms 'record linkage' and 'record linkage methodology' were used as input into Google Scholar, with results examined by year. Only papers that discussed the methodology of record linkage were counted and further classified – papers discussing the details of a specific linkage were excluded. The search for relevant papers per year was halted when results no longer appeared fruitful. For each year and search term, a minimum of 200 papers were examined.

The determination of appropriate classifications and the act of classification of individual papers was carried out entirely at the author's discretion. Given the diverse nature of record linkage, there are undoubtedly a number of relevant publications not found through our literature search; the selection of publications identified is hopefully representative of the entirety of the record linkage literature.

preserving record linkage; there were on average nine publications per year on this issue after 2008, as compared with a total of nine in the five years prior.

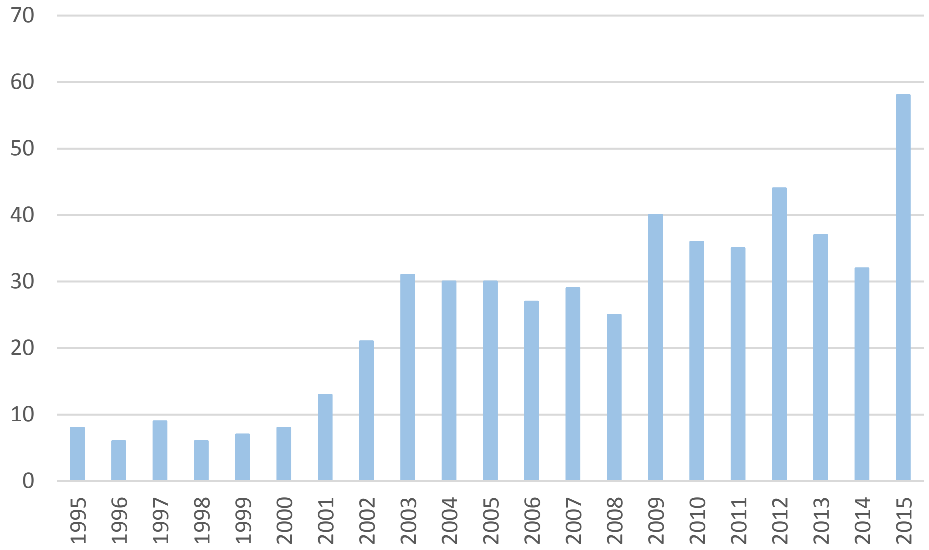


Figure 4: Number of record linkage methodology publications by year

Topic common within the literature are listed in Table 1. Some key trends are visible. Methods for improving privacy protection and linkage quality remain at the forefront of record linkage research, with a lesser focus on methods to improve the efficiency of record linkage. The presentation of novel linkage methods made up over a fifth of all papers on record linkage methodology.

Table 1: Common subject matter in 533 papers on record linkage published 1995-2015

Topic	*Number of papers*
General review	110
Methods for Improving Linkage Quality	**92**
String comparison methods	36
Data cleaning methods	18
Comparing non-western names	7
Estimating linkage parameters	17
Lack of independence of variables	5
Extensions to probabilistic linkage methods	14
Grouping	10
New linkage methods	108
Comparing and evaluating linkage methods	29

Machine learning approaches	43
Entity resolution	43
Methods for Improving Speed	**48**
Blocking	38
Parallel processing	7
Methods for Improving Privacy	**89**
Presentation of novel privacy preserving protocols	57
Evaluation of privacy preserving protocol	14
Presentation of record linkage software	36
Evaluation of record linkage software	9
Effective of linkage quality on research outcomes	35
Bayesian methods for record linkage	24
Automating linkage functions	24
Methods for linkage quality estimation	10
Synthetic data generation	6

The major topics identified in Table 1 are explored in more detail below.

Methods for improving linkage quality

Research has been conducted on all aspects of the linkage process, including data cleaning, matching, grouping, and post-linkage quality evaluation. These improvements typically aim to improve overall quality, but can also involve improvements to automate tasks that are typically manual and time-consuming.

Data cleaning involves the transformation of linkage variables prior to linkage to improve linkage quality. Several papers have aimed to extend the standard data cleaning methods; these include the development of modern algorithms for address and name standardisation (the splitting of these fields into their component parts) [83], and the development of robust systems to perform manual lookups (as well as other data transformations based on particular values) [84]. As the determination of rules for data cleaning is highly manual, there has been focus on automating these procedures. These typically involve use of unsupervised machine learning approaches, including the automatic development of lookup tables (for instance, nickname lookup tables, which contain a diminutive name along with their more common version) and the automated development of data cleaning rules from the data itself [85-87].

Despite widespread use, there has been little evaluation of the usefulness of data cleaning. One evaluation involved the use of lookup tables for names. It showed that results are highly dataset dependent [88], with the authors suggesting further testing is required to determine the overall effect of data cleaning on record linkage quality. As many data cleaning techniques are heavily manual and thus expensive and time consuming, any improvement in linkage quality must be weighed against these costs.

The use of different comparison types to account for uncertainty is common within probabilistic record linkage. As well as simple exact matching, string similarity techniques are regularly used on alphabetic fields to tolerate error. For instance, the values *Kate* and *Katie* would receive a high score using string similarity measures, but would be deemed incorrect using exact matching. The use of string similarity measures typically results in a large increase in linkage quality [89, 90], and is now considered standard practice. There have been several evaluations of string comparison techniques [89, 91-94], with the Jaro-Winkler string comparison algorithm emerging as the gold standard [89, 94]. New comparison techniques, aimed at particular problem sets, are still regularly developed [95-97].

Several works have explored additional string comparison techniques with parameters which adapt based on the nature of the data in question [98-101]. This approach achieves high quality, although it requires the use of a training dataset of known correct and incorrect pairs of fields. This requirement adds a layer of manual complexity which may limit its adoption.

The development of privacy preserving record linkage (discussed later in this Chapter and also in Chapter 5) has also resulted in a focus on privacy preserving string similarity measures. While privacy preserving options for standard measures such as the Jaro- Winkler comparator have been explored [102], q-gram based methods have been shown to be simple to implement in a privacy preserving setting [14], and achieve quality similar to standard un-encrypted techniques [94].

The comparison of non-Western names has received some attention in the literature. A number of phonetic encoding techniques such as Soundex [103] and NYSIIS [104] are used in record linkage. Phonetic encoding techniques aim to remove variation so that words which *sound* the same receive the same representation. The commonly used phonetic encoding techniques were originally developed for Western names, and most similarity measures have only been evaluated on Western names [105]. These methods will not necessarily provide accurate results for non-Western names. Alternate methods have been developed for use in non-English speaking countries [106-108] and in Western countries which now have higher levels of immigrations and thus a preponderance of non-Western names [109, 110]. Research has shown record linkage for immigrant groups in Western nations to be of poorer quality [7, 111]; as such, these developments are valuable.

Estimating linkage parameters

While there are a variety of linkage methods in use, all require some level of parameters as input into the linkage process. The process of determining these parameters is not always straightforward. For probabilistic record linkage, parameters are typically estimated by manual review (a time consuming process) or by re-using the parameters that have been used previously [112]. An alternate method suggested in the literature is to generate 'synthetic' data; made up records designed to imitate real administrative data, with the same types of error. This data could be generated with a truth set or 'answer sheet', identifying which records actually do belong to the same individual. The synthetic data can be linked using the estimated parameter settings, with the answer sheet used to evaluate their effectiveness [113] (with the assumption that these settings will also be valid on real data). Synthetic data can also be used as training data for estimating linkage parameters in linkage approaches which utilise machine learning, thereby removing the need to create a training set through manual review (machine learning approaches are described below) [114]. The EM algorithm has been used to estimate field weights for probabilistic linkage [112], achieving quite accurate results in many circumstances [115, 116]. The advantage of this algorithm is that it automates the estimation of parameters. However, there are instances where the EM model fails to converge, and research has looked at ways of ensuring the robustness of this method [117]. Methods for estimating the threshold setting for probabilistic record linkage, also often carried out manually, have been explored [116, 118]. Parameter estimation for other linkage techniques have been explored in the literature [119, 120]; as each linkage method has different parameters, each requires its own estimation method.

The accurate estimation of linkage parameters without manual intervention has become especially important in more recent times. Automation of record linkage procedures is increasingly important as demand for linkage services increases, and dataset numbers and sizes increase. Furthermore, the development of privacy preserving record linkage *requires* non-manual parameter estimation procedures, as manual review is not possible where input data is encrypted (this is discussed further in Chapter 5).

Probabilistic linkage methods appear capable of full automation when utilising the EM algorithm [116]. Future studies will be required to determine how accurate this fully automated method remains over time.

Extensions to probabilistic record linkage

As the most common form of record linkage, much research has focused on extending and enhancing the probabilistic method.

One of the key assumptions of probabilistic record linkage is the independence of variables for non-matching records; that is, having the same value of one field does not increase the likelihood of having the same value of a second field [121]. However this assumption clearly does not hold in typical record linkage; for instance, a person's first name provides information on their gender, and individuals with the same address are more likely to have the same surname. Research suggests this has the ability to bias results in certain circumstances [64]. Regression based methods have been used successfully to adjust for conditional dependencies [122, 123]. These methods improve linkage quality where the proportion of correct matches is high, or the level of conditional dependence is high [122].

The use of string comparison techniques [124] instead of exact matching has required the development of methods to determine appropriate settings [90, 125]. The use of weights based on individual field values (i.e. weighting a match on a common surname like 'SMITH' lower than a match on an unusual surname) also requires methods to determine appropriate parameter settings [126], although there has been some question as to the effectiveness of this technique, as evaluations have shown limited improvement [65, 127].

Grouping

The matching process generates a set of record-pairs; these are pairs of records thought to belong to the same individual. It is the grouping process which amalgamates these pairs into a more practical form - a linkage map listing which record belongs to which individual. The most common grouping strategy, known as *transitive closure* or merge-based grouping, amalgamates all record pairs above an accepted threshold, with all connected records classified as belonging to the same individual.

Several other alternate approaches to grouping have been put forward in the literature. Many of these alternate grouping methods utilise information about the relationships between all records to decide whether to accept or reject a pair [128-130]; however these techniques have not received any follow-up validation. An additional challenge occurs when each dataset being linked has pre-defined *groups* of records which belong to the same person; in this case, the grouping algorithm may wish to take into account the pairs formed between all members of one group with all members of another [131]. Methods have also been suggesting for ongoing record linkage, where incoming data is linked to a repository of previously created. These grouping methods aim to ensure that groups within the repository are not merged together [132, 133].

While matching nearly always occurs on a pairwise basis, full pairwise comparison may not be efficient enough for very large linkage involving hundreds of millions or billions of records. An alternate grouping algorithm reduces the number of comparisons by only conducting pair comparisons where records have not already been grouped together [134]. This method appears superior to the standard transitive closure methods, with improved efficiency but no change in results.

Linkage methods

Probabilistic record linkage, as well as deterministic rules based linkage, have been the two most common methods for matching. However, research into alternate methods for record linkage has appeared regularly in the record linkage literature. The two main techniques that have received focus have been record linkage methods utilising machine learning algorithms, and record linkage methods focusing on entity resolution (both described further below). A range of other approaches have been proposed, including statistical approaches including mixture models [135] and Bayesian methods [136-138], as well as approaches from the database literature [139, 140].

Machine learning

Machine learning algorithms form the basis of modern robotics and internet search amongst numerous other fields. Their defining feature is that they learn based on the supplied data. *Supervised* machine learning requires a 'training dataset' for which the answers are known. The algorithm first uses the training set to learn how to correctly carry out the task, before it is applied to data without an answer sheet. *Unsupervised* machine learning approaches refer to the class of algorithms which do not require a training dataset, but can be applied directly to the data at hand.

A large number of machine learning techniques have been suggested for record linkage. Supervised techniques include decision tree methods [114], support vector machines [141], maximum entropy modelling [142], the K nearest neighbour algorithm [143], random forests [144], the Electre Tre method [145], classification rule learning [146] and neural networks [147]. Unsupervised, or semi-supervised techniques include methods for support vector machines [148], K means clustering [149], cluster based decision models [150], bagging, bumping and Multiview methods [151], hierarchical graphical models [152] and an active learning process, which attempts to use provided secondary datasets as sources of additional information [153].

The machine learning record linkage methods presented within the literature have generally not been evaluated against other approaches (including [114, 141, 142, 144-147, 151, 153]). Support vector machines have showed promising results for certain linkage scenarios, compared to traditional probabilistic linkage [154]. One evaluation of machine learning methods suggests the best techniques are dataset dependent, but show that support vector machines and naïve Bayesian classifiers perform well [155]. Given the paucity of evaluations, there is little knowledge about which of these methods offer improvements over the standard probabilistic approach. As such, it is not surprising that most linkage software packages do not implement machine learning approaches (ChoiceMaker being the key counterexample [142]). These methods have also received little follow up within the literature; rather than building on previously established techniques, researchers have instead created their own approaches. Work by Christen on support vector machines [148] is one of the few examples of work building on previous approaches [149].

One of the key issues of the machine learning approach is the requirement for training datasets; these are necessary in supervised learning approaches for the procedures to internally set parameters, but generally require extensive manual effort to create. Several papers have focused on methods for automating some or all of this process. The use of synthetic data (generated with answers) as a training set has been suggested [113, 149], as have fully automated methods to determine test data [156] and methods which select only those training examples which will have the greatest impact, lowering the burden of manual review [157, 158]. Such developments will likely improve the adoption of machine learning approaches.

Entity Resolution

Entity resolution seeks to determine entities (i.e. the same 'things') between and within datasets. Entity resolution can be considered a broader field within which record linkage is contained. While in record linkage each record pertains to a single entity (typically an individual), the field of entity resolution does not require this assumption. A simple example of entity resolution would be to extract all the authors, papers and journals from a corpus of references. Here each record consists not of a single individual, but of a number of authors, a title, and a journal. The aim is to map these entities to a created canonical listing of authors, titles and journals.

Approaches to entity resolution are often sub-domain specific, with each approach aiming to make use of additional information unique to that set of problems. This includes approaches for online comparison shopping [159], personal information management [160], bibliographic data [161], extracting entity information from free text [162], identifying individuals based on behaviour, such as that found in transactional logs [163] and identifying individuals from social network information (where records of people with the same friends are more likely to be the same person) [164]. Several generic approaches to entity resolution have also been presented [165, 166].

There appears to be a disconnect between the entity resolution literature, mainly found in the computer science domain, and the record linkage literature, closer to the statistical and health domains. A number of entity resolution papers essentially deal with record linkage problems, without much reference to the record linkage literature [167-170]. Similarly, the record linkage literature has essentially ignored any work on entity resolution. Much could be gained by the cross-pollination of ideas from both of these sub-areas.

Comparing Linkage Methods

There have been relatively few comparisons of record linkage methods in the literature. There have been several comparison of deterministic (rules-based) record linkage and probabilistic record linkage, with results tending to favour probabilistic methods [171-174], although some find little difference between the methods [175], and others suggest each are useful for specific problems [25].

There are several challenges to comparing linkage methods. Firstly, record linkage datasets tend not to be publically available or shareable, meaning that different evaluations are carried out on different datasets, and new methods cannot be directly compared to the results of previous evaluations. The development and use of realistic synthetic data goes some way to solving this issue, as this data can be created with a truth set and shared [176]. A second and more difficult issue lies with determining the appropriate parameters when evaluating a linkage. Record linkage methods typically have a large numbers of parameters – for probabilistic record linkage alone there is the choice of the number of blocks and their composition, the comparison fields, the comparison methods, and the agreement and disagreement weight given for each field, as well as the final threshold method. Poor parameter selection will produce poor quality linkages, irrespective of the method used. This provides a challenge when evaluating the comparisons found in the literature. One solution to this problem is to adopt a particular method for determining parameters for each linkage method; the research question then becomes 'Is this linkage method, along with this method for determining its parameters, better or worse than this other linkage method, along with *this other* method for determining parameters?' By placing these assumptions at the forefront of the research question, readers can better assess the claims of an evaluation.

A final challenge is in determining the sensitivity of a linkage method to changes in parameters. Linkage methods need to perform similarly on different datasets, and not be susceptible to large changes in linkage quality due to small changes in linkage parameters. Such methods would be unlikely to perform well in realistic scenarios where estimating parameters may not be straightforward. This vital aspect has generally been overlooked in evaluations.

Linkage quality and research outcomes

Maximising linkage quality is not an end in itself; for health record linkage, the ultimate aim is to ensure reliable *research outcomes*. Several papers have observed the large effect that poor linkage quality can have on research outcomes [177, 178]; however knowledge of how errors (and different types of errors) directly affect particular methods of analysis is limited. False positive errors in one-to-one linkage of registry and death information have been shown to underestimate survival, while false negative errors will overestimate survival [179]. The relationship between these types of errors and more complex analyses is unknown. Gaining a greater understanding of the impact of false positives and negatives on particular methods of analysis, including whether either of these error types play a greater role in biasing outcomes, is an important research goal.

While not strictly an error of linkage quality, one study (*Publication 2 of this thesis)* [60] has measured the proportion of health records which occur outside a state jurisdiction (record linkage in Australia is typically conducted within states). Between 3% and 5% of individuals residing in one state were found to have hospital records in another state; as such, any state-based study of these individuals will necessarily be missing hospital records. The effect that this has in biasing research outcomes is not known, however.

A second related research question is to determine how a researcher can adjust or control for these errors, given their likely existence. Researchers currently make the assumption that linkage contains no errors, and do not attempt to adjust for linkage error. This may not be appropriate or correct, as linkage error is known to exist to some extent, and there is evidence that some subpopulations are more prone to linkage error than others, such as women [7, 180], the elderly [7], ethnic minorities [7, 111], certain geographic areas (from recording differences in specific localities) and those from lower socioeconomic groups [181].

Several methods for adjusting for linkage errors have been proposed within the literature. A team led by Chambers has presented a series of methods for different variations of one-to-one record linkage (linkage of two datasets, where each dataset contains at most one record per person). These methods include an estimation of linkage errors as an additional factor within regression analysis [182-184]. Several additional methods for including linkage error within regression analysis utilising Bayesian statistics have also been proposed [119, 185] A second approach, led by Goldstein, has been to utilise *all* record-pair associations, along with their confidence, in one-to-one matching rather than only the highest [186, 187]. This approach introduces the linkage error variation into the estimation of clinical variables, which are then used in statistical analysis. There has been limited empirical evaluation of these methods to date, and no comparison. The separation of personal identifiers from clinical information has made this research difficult to carry out in practice. Further work is required in this area, especially to extend this work to include linkages which involve datasets containing multiple records per person.

Methods for improving speed

The comparison of every record in one dataset to every record in another is too time-consuming to be practical. Blocking techniques are used to reduce the comparison space. Typically a set of *blocking variables* is chosen; only record-pairs which have exactly the same values of these variables will be compared further. Several sets of blocking variables are used, with record-pairs compared if and only if they have exactly the same values on one of these sets [10].

The development of new, improved blocking techniques, which compare fewer record-pairs without compromising record linkage quality, has received much attention. New techniques include blocking using suffix-arrays [62, 188], sorted neighbourhood methods [189, 190], canopy clustering [191], locality sensitive hashing [61] and q-grams [22]. Several reviews of blocking methods can be found in the literature [22, 192]. The adoption of more efficient blocking techniques is likely to become important with the linkage of extremely large datasets (100 million to billions of records) where speed may become a factor, or in specific linkage scenarios where speed is vital, such as real time linkage. However for typical health record linkage, where datasets are typically in the tens of millions of records, the time taken for matching is only a small part of the overall processing time [60]. The traditional blocking approach has been considered fast enough, and these alternate methods do not appear to have received adoption.

A second focus of research has examined automating choices of blocking variables [193-198] as well as comparison variables [199]. These techniques are most useful for linkages involving unusual variables or unusually large datasets. Standard record linkage typically involves the same variables (names, dates of birth, gender and address information); the use of default blocking criteria and comparison variables has likely been robust enough for automated record linkage.

A final aspect of research into blocking has looked at privacy preserving blocking, which seeks to provide ways to cluster records for comparison without revealing any identifying information. Simple methods involve the use of hashed (encoded) blocks [200]; however, in this case, block size may reveal a small amount of information. Some alternate methods have been proposed including the use of multibit trees [63, 201], clustering methods [202, 203], locality sensitive hashing [204] and reference datasets [205]. This is currently an active area of research.

Methods for improving privacy protection

The advancement of methods to further protect the privacy of individuals in record linkage remains a priority[2]. Privacy preserving record linkage involves carrying out record linkage on encrypted or encoded data; in this process, the linkage unit has no access to full personal identifiers, but only to some form of encoded information. As no personally identifying information is released by the data custodian, the privacy risk involved is much lower.

Privacy preserving record linkage is a popular area of current research, with an increasing number of protocols being proposed, each with different aims, approaches, and applicable scenarios. A comprehensive review of these methods has been undertaken [13]. This review highlights that most techniques are applicable to the linkage of two datasets only, and that privacy protections offered

[2] The requirement for privacy in record linkage, its legal implications, and common operational methods for ensuring privacy are discussed in Chapter 5. This section focuses only on new developments in the literature regarding privacy, namely privacy preserving record linkage.

by these methods require “honest parties”; that is, it requires that there is no collusion between a linkage unit and a data custodian to learn about a second custodian’s data.

Privacy preserving protocols presented in the literature utilise either a two or three party model [13]. Under a two-party protocol, only those organisations that hold data participate in the linkage – there is no independent third party acting as linkage unit. This differs from almost all current models of record linkage used today, such as those in the UK [206], Canada [207] and Australia [59], which utilise third parties. These protocols typically require a greater amount of communication than three-party protocols [14].

Privacy preserving protocols also differ in the level of privacy protection they provide. The least privacy preserving techniques simply amalgamate parts of a person’s identifiers into a single variable [15]. Although not quite as identifiable as no privacy preserving method at all, it is a trivial exercise for a person undertaking linkage or a malicious individual to determine whether an individual exists within a database of these codes.

Another class of privacy techniques encrypt data so that those with access cannot learn any information directly from the encrypted values; however, these encrypted values can be vulnerable to frequency attacks [16]. For instance, a protocol may encrypt fields separately, all in a particular way – every instance of ‘John’ will have the same encrypted value. As John is the most common name, it is possible to count the frequency of encrypted values to work out which value corresponds to this name. These protocols are significantly more complex to break, and typically only some percentage of information can be revealed. Hashing (a type of irreversible data transformation) is the most common technique used in these circumstances [13, 14, 208, 209].

A final class of privacy techniques encrypts data in such a way that it is not possible to learn any information about individuals. Such methods utilise strong cryptographic techniques (the same as those used to ensure security in other domains such as finance and the internet) for which there are no known attacks. Unfortunately, while protocols using these techniques exist [210-212], they are currently impractical for all but the smallest applications of record linkage [213].

Another important difference in these record linkage privacy protocols is the method of matching. Protocols either perform an exact match on a particular set of identifiers, or perform similarity matches on particular fields. As discussed earlier, approximate similarity matching is vital to obtain high linkage quality. Several methods have been suggested to allow privacy preserving string comparisons, including embedding identifiers within a metric space [214, 215], using public reference datasets [216, 217], and hash based methods utilising Bloom filters [14].

A final difference found amongst these protocols is the extent to which they have been used in practice. Almost all the privacy preserving linkage methods that have been used in operational settings have involved an exact match on particular attributes of a dataset [218, 219], which are generally hashed to ensure privacy [208, 220, 221]. These methods are efficient, and relatively simple to implement, although the resulting linkage quality is sub-optimal [222].

In line with other aspects of the record linkage literature, while there have been a large number of proposed protocols, very few have received any strong empirical evaluation, especially using large real administrative datasets. Without this, it is unlikely that these methods will be adopted by operational linkage units. In addition, few methods have received follow up in the literature; a notable exception is the work of Schnell, who first proposed the use of Bloom filters for approximate privacy preserving linkage [14]; this seminal work has resulted in the development of a number of other protocols [223-225], as well as extensions and comparisons of the method [94, 226-229].

Based on the available evidence, the privacy preserving method using Bloom filters appears to be the most likely of the numerous privacy preserving protocols to receive adoption by operational linkage units. Evaluations have shown equal linkage quality and similar efficiency to traditional unencrypted probabilistic linkage [200]. The security of the protocol has been investigated in detail [16, 230, 231], with numerous refinements made to strengthen it [229].

Conclusions

From examining the literature on methods for record linkage, there are several identifiable trends.

The record linkage literature has predominantly focussed on *creation of methods* rather than *evaluation of methods*. This can clearly be seen in the development of machine learning methods for record linkage, and the development of privacy preserving protocols. In both of these areas, a very large number of new methods have been presented, all of which attempt to solve the same problem. Notwithstanding their novelty, few of the methods have been evaluated to determine whether they offer significant improvement over current approaches. As a result, many of these methods are essentially orphaned, receiving no further evaluation or follow-up by the research community.

Further systematic testing of new linkage methods is necessary if new methods are to receive adoption. Rigorous testing and comparison of machine learning methods against currently used linkage approaches appears to be an important next step if such methods are to be adopted in practice. These methods would need to be evaluated not only for quality and efficiency, but also in terms of the difficulty of determining appropriate parameters (or creating appropriate training data) as well as the methods sensitivity to changes in parameters.

The privacy preserving linkage literature appears to be better placed, with research coalescing around the Bloom filter method. Further research is required to determine how to correctly estimate parameters in a privacy preserving context, and to ensure linkage quality remains high in light of security improvements to the protocol. The development of further protocols which provide higher security at a similar level of accuracy and speed would also be an important development.

Many methods which have not received evaluation may still have merit. For instance, it seems unlikely that the probabilistic record linkage method, now over 50 years old, would outperform state of the art machine learning methods, a huge area of advancing research over the last 20 years. However without detailed, rigorous effort to test these methods, practitioners have little choice but to continue with known and well-tested methodology.

The reason for this focus on creation rather than evaluation of methods is not clear, but several causes are suspected. It may be partly explained by a type of publication bias, where academic journals are possibly more reluctant to publish evaluations of previously published methods, preferring 'original research' articles. A similar publication bias has been noted in the literature for replication studies [232]. There is no requirement by reviewers for detailed evaluations of new methods and protocols, as can be seen by the number of papers without these. High quality evaluation itself is difficult, requiring access to large real-world datasets with "known" answers. Such datasets are difficult to source and almost impossible to share, making comparative evaluation difficult [176]. The use of synthetic data with known answers may be one way around this issue [176]. Carrying out evaluations will also occasionally determine that the proposed novel methods are not actually better than previous methods; indeed, there is evidence to suggests that null results are much harder to publish [233].

There may also be a disconnect between theory and practice. Research into record linkage methodology typically occurs within academia, while most record linkage units are located within government. The communication between these two groups could be improved; academia has long been described as an 'ivory tower', disconnected from practical concerns [234]. It is possible that academic researchers do not realise the necessity of evaluations of their methodology for it to be used in practice. Record linkage practitioners usually reside in government, which has traditionally been considered risk averse [1] and hostile to innovation [235]. Record linkage units require a certain level of confidence in a new technique before they invest in evaluation or implementation. Greater collaboration between academics and practitioners would be one of the most important ways to improve adoption of new record linkage methods in practice.

Despite these misgivings, it is clear that great strides have been made in record linkage methodology. Dataset sizes, and dataset numbers have grown dramatically and methods to handle this increase have kept pace. The literature is replete with methods to handle further orders of magnitude increases in dataset sizes [22, 134]. Probabilistic record linkage methods have been further refined, and while new linkage methods have not received proper evaluation, once this occurs there will undoubtedly be methods which further extend the state of the art. Privacy-preserving methods appear close to ready for use in operational settings, with the adoption of these methods likely to lead to greater dataset access for researchers. With a greater focus on evaluation and collaboration with practitioners, these significant developments will undoubtedly improve linkage quality and access to datasets.

Chapter 4

Methods for improving quality

Research Output

Supporting Publications

8. Boyd, J. H., Ferrante, A. M., Irvine, K., Smith, M., Moore, E., Brown, A. P., **Randall, S. M.** (2016). **Understanding the origins of record linkage errors and how they affect research outcomes.** *Australian and New Zealand Journal of Public Health, In Press*

9. Boyd, J. H., Guiver, T., **Randall, S. M.**, Ferrante, A. M., Semmens, J. B., Anderson, P., & Dickinson, T. (2016). **A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects.** *Methods of information in medicine*, *55*(3), 276-283.

Key Publications

10. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). **The effect of data cleaning on record linkage quality**. *BMC medical informatics and decision making, 13*(1), 64.

11. **Randall, S. M.**, Boyd, J. H., Ferrante, A. M., Brown, A. P., Semmens, J. B. (2015). **Grouping methods for ongoing record linkage**. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

12. **Randall, S. M.**, Boyd, J. H., Ferrante, A. M., Bauer, J. K., & Semmens, J. B. (2014). **Use of graph theory measures to identify errors in record linkage.** *Computer methods and programs in biomedicine, 115*(2), 55-63.

4.1. Measuring and reporting linkage quality

Achieving high linkage quality is essential to ensure the validity of any research utilising linked data. Linkage quality here refers to the proportion of errors found within a linkage. There are two possible types of errors. Firstly there are false positives, where two records are designated as belonging to the same person, whereas in reality, they belong to two different people. Secondly, there are false negatives, where two records are designated as belonging to two separate people when in truth they belong to the same individual.

There is often limited reporting by linkage units regarding the quality of their linkage to researchers, and it is not clear that researchers have much understanding regarding linkage quality. Linkage errors are not evenly distributed throughout the population, but tend to cluster around specific demographics; women [7], ethnic minorities [7], and those from lower socioeconomic groups [181], providing greater potential to bias results in these populations. Further research in this area is needed. These issues are explored in *Publication 8: Understanding the origins of record linkage errors and how they affect research outcomes [236].*

The two types of errors found in record linkage (false positives and false negatives) are typically measured as precision (the proportion of matches found that were correct) and recall (the proportion of correct matches that were found) [237]. A linkage with a high precision will have few false positives, while a linkage with high recall will have few false negatives. The F-measure of a linkage is the harmonic mean of precision and recall; it provides a single value with which we can compare results.

To be able to calculate the precision and recall of a linkage, we need to know in advance whether individual record-pairs have been correctly identified or not. However in practice, this information is not known. One proposed solution to this problem involves sampling record-pairs at different threshold scores, and manually reviewing these samples. This new method is presented and evaluated in *Publication 9: A simple sampling method for estimating the accuracy of large scale record linkage projects [238].* This sampling based method appears highly accurate.

While there is often little reporting of overall linkage quality to researchers, efforts to optimise linkage quality play a large role in linkage operations. Current methods can involve manual review of record-pairs to ensure quality [8, 9], a costly and time consuming process. By determining best practices for record linkage operations with regard to linkage quality, and by developing new methods to optimise quality and reduce the manual processing burden, this thesis aims to allow higher linkage quality to be achieved more quickly and at lower cost. In order to tackle this issue, this thesis has focused on specific constituent parts of the record linkage process.

4.2. The linkage process

The record linkage process is presented in Figure 5.

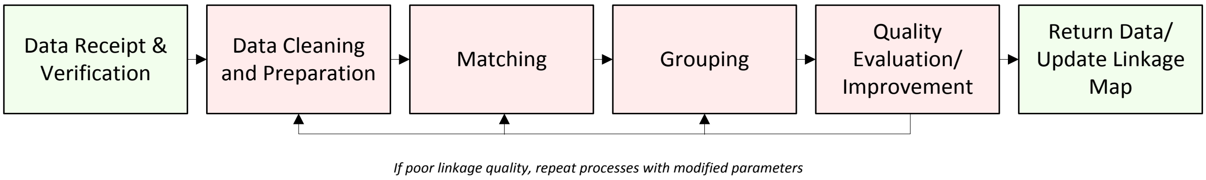

Data Receipt & Verification
Data Cleaning and Preparation
Matching
Grouping
Quality Evaluation/ Improvement
Return Data/ Update Linkage Map
If poor linkage quality, repeat processes with modified parameters

Figure 5: The record linkage process

Once data is received, it is cleaned and standardised, a process where individual fields of the dataset are modified, replaced or removed in order to improve the underlying data quality. The matching process follows; individual records are compared, typically in a pairwise fashion, to determine whether the two records belong to the same person. This process results in a list of pairs of records believed to belong to the same person.

The *grouping* process then merges these record-pairs into its final format; a listing of each record and the individual to which it belongs (a linkage map).

Once the linkage map is created, methods are used to measure and improve linkage quality. If linkage quality is low, one approach is to repeat previous steps of the linkage process with alternate parameter settings. Manual methods such as clerical review are often used by linkage units to improve quality; however these are very slow and thus expensive. Manual methods also do not scale easily; in a large linkage of many millions of individuals, the manual resolution of an incorrect linkage of one individual will have a miniscule if any effect on overall quality.

Each of these components is discussed in further detail below.

Data cleaning and preparation

Data cleaning involves the transformation of the information received for linkage into a standardised format that is most appropriate for matching. A variety of techniques are used to prepare data for linkage; these can be found in various linkage software packages and in the literature [10, 239, 240].

Data of higher quality (containing fewer errors and ambiguities) will result in fewer mistakes in linkage, and thus return higher linkage quality. Improvements in data quality should therefore translate into improvements in linkage quality. Despite this widely held view, there has been surprisingly little empirical evaluation of the effect of data cleaning on overall record linkage quality, or of which particular data cleaning techniques are most useful.

Publication 10: The effect of data cleaning on record linkage quality [241] outlines methods for data cleaning and their prevalence, and evaluates the impact of these methods on linkage quality, using both synthetic and real administrative data. The results suggest that rather than lead to noticeable improvements in linkage quality, heavy data cleaning can reduce the overall quality of linkage.

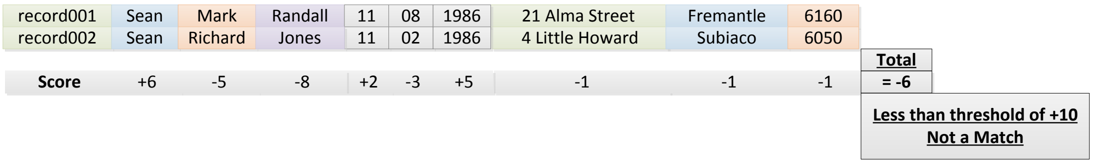
Given these findings, the allocation of significant resources to data cleaning (up to 75% of time by some estimates [10]) may not be warranted. These results also suggests that the use of particular techniques, (such as name lookups) should rarely, if ever, be used as they are likely to lead to an overall decrease in linkage quality. Further testing on additional datasets will shed more light on exactly how generalizable these results are.

Matching

The matching process takes as input the cleaned and prepared datasets, and in turn produces a list of record-pairs, where each record in a pair is considered to belong to the same individual [10]. Matching can occur between two or more files, or within a single file (known as *de-duplication*); the process is essentially the same in either scenario. A simple approach to this problem is the use of deterministic matching [10]. In deterministic matching, a combination of personal identifiers is chosen such that if two records have all of these values in common, they are designated to belong to the same individual. Each single combination of personal identifiers is known as a rule. In practice, a series of rules are typically used, in order to allow some tolerance of missing, incorrect or changing values. While deterministic linkage is intuitive and quick, it is less versatile and achieves lower linkage quality than probabilistic record linkage [174], the most common method for record linkage.

Probabilistic linkage

Probabilistic linkage, so called because it uses conditional probabilities to compute likelihoods, presents a more formal and less intuitive approach to linkage. In this approach, records are compared on a pairwise basis. A comparison of two records involves comparing all the individual field. Each field comparison results in a score based on specific weights assigned to that field. These scores are summed up for the pair comparison, and if this summed score is over a specific threshold, the two records are designated a match (see Figure 3) [121].



record001	Sean	Mark	Randall	11	08	1986	21 Alma Street	Fremantle	6160	
record002	Sean	Richard	Jones	11	02	1986	4 Little Howard	Subiaco	6050	Total
Score	+6	-5	-8	+2	-3	+5	-1	-1	-1	= -6

Less than threshold of +10
Not a Match

Figure 6: An example of a record-pair comparison in probabilistic linkage.

Probabilistic linkage removes the difficulties of creating large sets of rules and determining the validity of each rule. However the process still requires a number of parameters to be accurately estimated to ensure high quality linkage.

In probabilistic linkage and many other approaches, the comparison of every record in one dataset to every record in another quickly becomes infeasible as dataset sizes increase. For example, a relatively small linkage of 1 million records to a second dataset of 1 million records would involve 1 trillion record pair comparisons. The vast majority of these potential comparisons would involve records that belong to different people. Techniques known as blocking are used to dramatically reduce this comparison space, while keeping all (or very nearly all) comparisons which have a possibility of belonging to the same individual [10]. The typical method of blocking is to choose a set of fields (*blocking variables);* only record-pairs which have exactly the same values of these variables will be compared further [10].

Grouping

The matching process results in a set of record-pairs; these are pairs of records thought to belong to the same individual. The grouping process amalgamates these pairs into a more usable form - a linkage map listing which records belongs to which individual.

The standard grouping strategy is *merge grouping* or transitive closure [242]. In this process, all record pairs above the accepted threshold are amalgamated, with all connected records classified as belonging to the same individual (see Figure 7). Indirect relationships can be formed between records, where despite not forming a record pair themselves, they are joined by an intermediate record which has formed a pair with both. This method is intuitive, and is most commonly used when de-duplicating a file or linking files together which each contain multiple records per person.

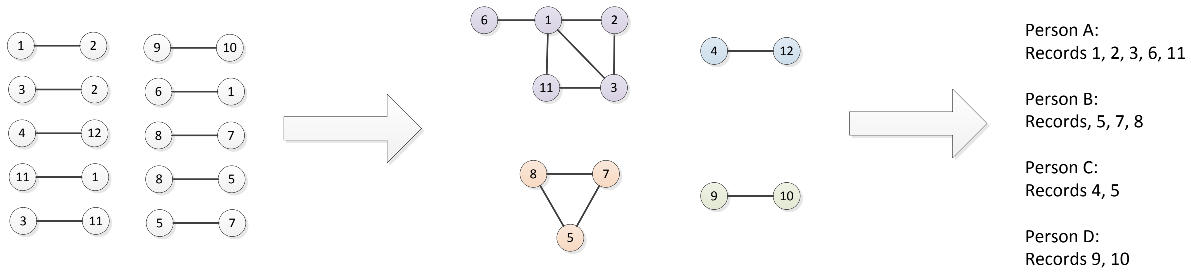


Figure 7: The merge grouping process

As linkage units grow, they typically move to an on-going *repository* model, where a single linkage map is created and maintained over time (see *Publication 1* for more detail on this approach). This linkage map is then used for multiple research projects. This approach can improve overall linkage quality, as the impact of quality intervention and manual assessment is not lost after each research project. As such, linkage units often have high confidence in their stored results.

An alternate grouping approach may be useful in scenarios where new data is being linked to a repository of previous linked data [132]. This alternate approach would ensure that the results of previous linkages are maintained, rather than potentially merge together groups of records when there is already evidence that they are separate people. Alternative methods of grouping have been proposed which would solve this problem [132], but few have been fully articulated, and fewer still have been formally evaluated.

Publication 11: Grouping methods for ongoing record linkage [243] outlines existing methods for grouping with a repository of previously linked data, and proposes an additional new method. The paper evaluates these methods using real administrative datasets. Results suggest that alternate 'best-link' approaches achieve substantially better quality. These methods should be recommended in scenarios where incoming is linked to an ongoing repository.

The grouping process produces a complete linkage map, listing each record with a corresponding person identifier. Before the linkage map can be used for its ultimate purpose (to extract health data for researchers), its quality must be evaluated. Additional techniques for improving quality can also be conducted.

Techniques for improving quality after linkage

The discovery of common types of errors through manual checking of results can often identify issues that can best be solved through altering the parameters of the linkage process and re-running these steps. This could include the modification of matching weights, editing the blocking parameters, or utilising additional cleaning techniques.

At record-pair scores close to the threshold cut-off, the likelihood of an incorrect classification (either a false positive or a false negative) is highest. Instead of using a single threshold score to determine which records are correct, some organisations use both a lower and upper thresholds score – those record-pairs that score in between these thresholds are manually reviewed [121]. This large-scale manual review process results in high linkage quality, but requires a significant amount of time and expense. Human manual review is slow, and for large datasets this method is infeasible. The exploration and establishment of additional techniques to reduce and remove the burden of large-scale manual review is an important area for development.

The creation of record-pairs and their amalgamation into a linkage map provides additional information unavailable during linkage that has potential utility for improving linkage quality. One source of additional information is in the structure of the pairwise-relationships formed through matching. These may convey additional information useful in identifying groups likely to contain false positive errors. If groups that are in error can be identified with high accuracy, corrective action can be more effectively targeted, thereby reducing the burden of clerical review.

An intuitive hypothesis is that groups of records which are more sparsely linked and held together may be more likely to contain false positive links, as compared with groups which are dense or fully saturated with pairs. The non-existent record-pairs in sparser groups likely had the opportunity to form a record-pair in linkage, but did not, potentially suggesting we should have a lower confidence in this group. Metrics from graph theory, the branch of mathematics which studies the structure of pairwise relations, could be used to identify these groups.

Publication 12: Use of graph theory measures to identify errors in record linkage [244] adopts several metrics from graph theory and applies these to linked groups of real world data. These metrics accurately identify groups containing errors, and present superior precision to traditional threshold setting methods. Results from this research provide the foundation for developing new methods for resolving these errors automatically.

Conclusion

Once the quality of the linkage is deemed satisfactory, the process is complete and the data is ready to be used by researchers.

Linkage processing is complex and is made up of a number of discrete parts. Each of these parts influences the overall linkage quality. The publications in this section have focused on particular record linkage processes, with an aim to either evaluate current practice with respect to its effect on linkage quality, or to develop new methods to improve overall linkage quality. This exploration of methods to improve linkage quality is by no means comprehensive or complete, but each development is significant in its own right. The methods presented here aim to reduce the burden of significant manual intervention. Improving linkage quality through these methods can both significantly decrease the cost of linkage operations, allowing more researchers access to data, and increase the confidence in research outcomes derived from linked data.

Publication 8

Boyd, J. H., Ferrante, A. M., Irvine, K., Smith, M., Moore, E., Brown, A. P., **Randall, S. M.** (2016). **Understanding the origins of record linkage errors and how they affect research outcomes.** *Australian and New Zealand Journal of Public Health, In Press*

Contribution:

SR supported the development of this paper, assisting with writing the first draft of the manuscript and contributing to the final version.

Publication 9

Boyd, J. H., Guiver, T., **Randall, S. M.**, Ferrante, A. M., Semmens, J. B., Anderson, P., & Dickinson, T. (2016). **A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects.** *Methods of information in medicine*, *55*(3), 276-283.

Contribution:

SR supported the development of this paper, collecting and analysing data, and contributing to the final version of the manuscript.

Publication 10

Randall, S. M., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). **The effect of data cleaning on record linkage quality**. *BMC medical informatics and decision making, 13*(1), 64.

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, performed all evaluations, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

BMC
Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

The effect of data cleaning on record linkage quality

Sean M Randall*, Anna M Ferrante, James H Boyd and James B Semmens

Abstract

Background: Within the field of record linkage, numerous data cleaning and standardisation techniques are employed to ensure the highest quality of links. While these facilities are common in record linkage software packages and are regularly deployed across record linkage units, little work has been published demonstrating the impact of data cleaning on linkage quality.

Methods: A range of cleaning techniques was applied to both a synthetically generated dataset and a large administrative dataset previously linked to a high standard. The effect of these changes on linkage quality was investigated using pairwise F-measure to determine quality.

Results: Data cleaning made little difference to the overall linkage quality, with heavy cleaning leading to a decrease in quality. Further examination showed that decreases in linkage quality were due to cleaning techniques typically reducing the variability – although correct records were now more likely to match, incorrect records were also more likely to match, and these incorrect matches outweighed the correct matches, reducing quality overall.

Conclusions: Data cleaning techniques have minimal effect on linkage quality. Care should be taken during the data cleaning process.

Keywords: Data cleaning, Data quality, Medical record linkage

Background

Record linkage in context

Record linkage is the process of bringing together data relating to the same individual from within or between datasets. This process is non-trivial when unique person based identifiers do not exist, and linkage is instead performed using probabilistic or other techniques that compare personally identifying information such as name and address, which may include error or change over time.

While record linkage is frequently performed in a business or administrative context to remove duplicate entries from person based datasets, it has also been widely used to enable health researchers to gain event based longitudinal information for entire populations. In Australia, research carried out using linked health data has led to numerous health policy changes [1,2], and the success of previous linkage efforts has led to the development of national linkage infrastructure [3].

Record linkage methodology

Approaches used in record linkage fall across a spectrum between deterministic and probabilistic methods. Deterministic linkage methods range from simple joins of datasets by a consistent entity identifier to sophisticated stepwise algorithmic linkage which includes additional information to allow variation between records that match i.e. it does not rely on an exact match of the entity identifier. Probabilistic methods, on the other hand, use various fields between data sets to calculate the odds that two records belong together [4]. These odds are represented as probability weights or scores which are calculated (summed) for each pair of records as they are compared. If the total score for a record pair is greater than a set matching threshold, then they are deemed to be a match – the records belong to the same person. The probabilistic approach allows for inconsistencies

* Correspondence: sean.randall@curtin.edu.au
Centre for Data Linkage, Curtin Health Innovation Research Institute, Curtin University, Perth, WA GPO U1987, Australia

between records with missing matches i.e. it has the capacity to link records with errors in the linking fields.

Several studies have demonstrated that probabilistic linkage techniques are more robust against errors, and result in better linkage quality than deterministic methods [5-7]. Probabilistic methods are also more adaptable when large amounts of data require linkage [8].

Data cleaning in record linkage

Irrespective of which linkage approach is being used, the linkage process is usually preceded by a data cleaning phase. Data cleaning (sometimes called standardisation or data cleansing) involves correcting, removing or in some way changing fields based on their values. These new values are assumed to improve data quality and thus be more useful in the linkage process.

There is evidence that improvements in the quality of the underlying data lead to improvements in the quality of the linkage process. For example, early studies of probabilistic linkage in health research demonstrated that greater amounts of personal identifying data greatly improved the accuracy of linkage results [9,10]. Studies have also shown that data items with more discriminating power lead to better linkage results [11,12].

In the absence of strongly identifying personal information, data cleaning has been recognised as one of the key ways to improve the quality of linkage [13]. The record linkage literature identifies data cleaning as one of the key steps in the linkage process [14-17], which can take up to 75% of the effort of record linkage itself [18].

Data cleaning techniques

A variety of data cleaning techniques are used in record linkage [18-20]. Some data cleaning techniques seek to increase the number of variables by splitting apart free text fields. Others seek to simply transform variables into a specific representation, without actually changing the information. Further techniques aim to change the information in the fields, either by removing invalid values, changing values, or imputing blank values. Based on a review of five institutions conducting linkage in Australia and eight linkage software packages [19], the following data cleaning techniques were identified.

Reformatting values

Data values can be simply changed to a new format without actually creating or removing information. This ensures that all data is in a common standard for comparison during linkage. For example, two datasets which store dates in a different format (such as '11/08/86' and '11th August 1986'), would need to be changed to a common format for comparison. No data is changed by this transformation, only the representation of the data. This technique is essential for ensuring matching fields can be compared [18].

Removing punctuation

Unusual characters and punctuation are typically removed from alphabetic variables. Names with spaces, hyphens or apostrophes may be more likely to be misrepresented, and removing these values can remove any differences between these values.

Removing alternative missing values and uninformative values

Datasets can often contain specially coded input values when no information is available – for instance '9999' for a missing postcode. Other datasets may contain information that is not useful to the linkage process - hospital admission records may contain 'Baby of Rachael' in a forename field, or 'NO FIXED ADDRESS' in an address field. These are commonly removed [18]. In traditional probabilistic linkage, two variables that agree on a value (for instance, both are marked 'UNKNOWN ADDRESS') will receive a positive score, which in this case, may be inappropriate. A comparison involving a missing or blank value will typically not result in any positive or negative score.

Phonetic encoding

By creating an encoding of the phonetic information encapsulated in an alphabetic variable (such as a surname) names that are recorded as different spellings but sound the same will be brought together. Phonetic encoding is a common technique in record linkage. Common encoding algorithms used in record linkage include Soundex [21], NYSIIS [22] and Metaphone [23]. NYSIIS has been used for record linkage in Canada [13], while in the Oxford Record Linkage Study the Soundex value of the NYSIIS code is used in their linkage [18].

Name and address standardisation

Name standardisation or name parsing is the process of breaking down a person's full name into its individual components. For instance, a name field with the entry 'Dr John Harry Williams' could be broken down into title, first name, middle name and last name, and these components could be individually compared.

Similarly, an address can be broken down into its constituents such as street number, street name and street type. By creating multiple variables in this way, small differences between records such as a different order may have less effect in bringing these records together. Typically the process of breaking the address into separate components has been carried out using a set of rules [24], but the application of statistical methods has also proved useful [25].

Nickname lookups

A nickname file, containing common nicknames and diminutive names for given names can be used to translate forenames to a common value. Using a nickname lookup, a person recorded as Bill on one dataset and William on another could be given the same first name, potentially bringing these records together [18].

Sex imputation

A record with a missing sex value can have this value imputed based on their first name. This requires a lookup table which equates common first names with sex.

Variable and field consistency

Records containing variables which are inconsistent can be edited to remove this inconsistency [20]. For instance, a record with suburb of Sydney and postcode of 6000 is inconsistent, as this is the incorrect postcode for this suburb. It is not often clear which variable to change in order to resolve this inconsistency.

Prevalence of data cleaning

These techniques encapsulate those found in linkage software packages or in use by dedicated linkage units in Australia during our environmental scan. All techniques listed here were either in use or under consideration by at least one institution conducting linkage in Australia, and all institutions asked used at least one of these techniques to clean their data.

A review of the data cleaning features found in linkage software packages can be found in Table 1. These linkage packages vary from enterprise level commercial packages (IBM's QualityStage [26]), smaller commercial packages (Linkage Wiz [27] and the now freely available Choicemaker [28]), free university developed software (Febrl [29], FRIL [30], The Link King [31]) and government developed software obtained for evaluation (LINKS [32], BigMatch [33]). Linkage engines are probabilistic (BigMatch, FRIL, Linkage Wiz, FEBRL) a combination of both rules based and probabilistic (LINKS, Link King) or using modern machine learning techniques (ChoiceMaker, FEBRL). Nearly all packages implement data cleaning as a set of functionality which the operator can choose to apply on specified variables in a dataset. In some packages (for instance, The Link King) data cleaning is performed as an automated part of linkage itself, with the operator having little manual control over the steps taken.

Data cleaning functionality in linkage software packages ranges from non-existent (BigMatch, LINKS) to comprehensive (Febrl. QualityStage, Linkage Wiz). Techniques available for reformatting variables typically include trimming, splitting and merging fields, classifying values, and reformatting dates.

Packages which remove specific values typically use a default invalid value list, which can then be added to by the user (for example Febrl, Link King, QualityStage, Linkage Wiz). Phonetic encoding algorithms available typically include Soundex at a minimum, with NYSIIS also common. Additional available techniques include 'backwards NYSIIS,' metaphone and double metaphone. The lack of data cleaning functionality in some packages tended to be the result of a design decision to split this functionality into a separate software package rather than a value judgement about its usefulness.

Advantages of data cleaning

In a record linkage context, the aim of data cleaning is to improve linkage quality [18,34]– that is, reduce the number of false positives (two records incorrectly identified as belonging to the same person) and false negatives (two records incorrectly identified as not belonging to the same person). Without data cleaning, many true matches would not be found, as the associated attributes would not be sufficiently similar [35].

Despite its widespread availability in linkage software packages, its use by numerous linkage groups, and its recognition as a key step in the record linkage process, the record linkage literature has not extensively explored data cleaning *in its own right*. Particular methods of cleaning data variables have been evaluated previously. Churches et al. [25] compared rule based methods of name and address standardisation to methods based on probabilistic models, finding more accurate address

Table 1 Availability of data cleaning functionality across a sample of linkage packages

	Linkage Wiz	Febrl	BigMatch	Link king	FRIL	LINKS	ChoiceMaker	QualityStage
Reformat values	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Remove punctuation	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Remove alt. missing values	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Phonetic encoding	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Name/Address Standardisation	Yes	Yes	No	No	No	No	Yes	Yes
Nickname lookup	Yes	Yes	No	Yes	No	No	No	Yes
Sex imputation	Yes	Yes	No	Yes	No	No	No	Yes

information when cleaned using probabilistic models. Wilson [36] compared phonetic algorithms and hand curated mappings on a genealogical database, finding the hand-curated mappings more appropriate for name matching. To our knowledge there has been no systematic investigation of the extent to which data cleaning improves linkage quality, or which techniques are most effective.

Objectives

Implicit in the data cleaning process is the assumption that data cleaning will improve linkage quality. However there is limited literature that has quantified the extent of improvement arising from data cleaning. Moreover, little is known about the relative effectiveness of various techniques. The current study attempts to answer these questions through a systematic investigation of the effect of data cleaning on linkage quality using two datasets – a ‘synthetic’ dataset and a large-scale ‘real world’ administrative dataset.

Since real world datasets for which the ‘answers’ are known are both difficult to source and virtually impossible to share, we opted to generate and use a synthetic dataset. The synthetic data files contain artificially created records that have characteristics that closely resemble the attributes of real world datasets. Such datasets are typically use in benchmarking or systems testing.

Methods

This study aimed to investigate both the overall combined effect of data cleaning, as well as the individual effects of specific data cleaning techniques. Firstly to investigate the overall quality, a highly cleaned, a minimally cleaned, and an uncleaned version of each of the two datasets was produced. These were each internally linked, with the resulting linkage quality measured. To investigate the effect of specific data cleaning techniques, the relative improvement of each transformation on the above datasets was measured and averaged Figure 1.

Datasets

The synthetically generated data set consisted of 400,000 records, containing multiple records belonging to the same person. The synthetic data was generated using an amended version of the FEBRL data generator [37]. As a first step, the generator creates a user specified number of original records. These are created randomly, based on frequency lookup tables. Duplicate records are created in a second step, based on the original records. Duplicate records are created by randomly selecting an original record, then randomly choosing the number of duplicates to be created from it, and then randomly introducing errors according to user-specified parameters. An additional probability distribution specifies how likely data items or attributes are selected for introducing errors (it is possible for data items to have no errors at all).

The synthetic data file was based on frequency distributions obtained from the Western Australian electoral roll. As voting is compulsory in Australia, the electoral rolls are highly representative of the population. To avoid the potential of identifying individuals from the electoral data, the frequency list was truncated so that frequency counts below five were excluded.

Each record in the dataset comprised the following data items: surname, first name, sex, date of birth and postcode. Records in each dataset were generated with errors typically found in administrative data. Ascertaining representative rates of different types of errors such as duplications, omissions, phonetic alterations and lexical errors involved abstracting errors manually from a number of real world datasets and extrapolating these to the artificial data. Real world errors were applied to the synthetic data using user-specified parameters which are

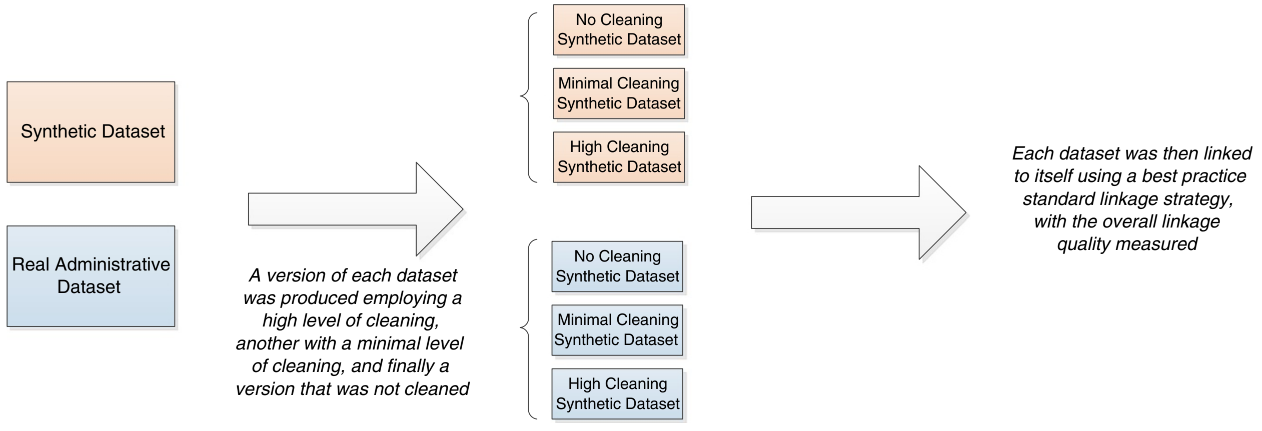

Synthetic Dataset
Real Administrative Dataset
A version of each dataset was produced employing a high level of cleaning, another with a minimal level of cleaning, and finally a version that was not cleaned
No Cleaning Synthetic Dataset
Minimal Cleaning Synthetic Dataset
High Cleaning Synthetic Dataset
No Cleaning Synthetic Dataset
Minimal Cleaning Synthetic Dataset
High Cleaning Synthetic Dataset
Each dataset was then linked to itself using a best practice standard linkage strategy, with the overall linkage quality measured

Figure 1 Road map for measuring overall linkage quality.

part of the Febrl data generator. Errors in the final dataset included the use of equivalent names, phonetic spellings, hyphenated names, first and last name reversals, change of surname, partial matches, typographical errors, incomplete or inaccurate addresses (postcode only) and change of address (postcode only). As Table 2 demonstrates, the synthetic datasets were highly representative of the source population.

This dataset had previously been used for an evaluation of linkage software [38]. An advantage to the use of synthetic datasets is that they are transportable, and so allow easier validation, and the 'answers' as to which records belong to the same person are available, unlike in real administrative data. This dataset is freely available (see Additional file 1).

Ten years of 'real world' hospital admissions data was sourced from one Australian state. This consisted of almost 7 million records. This dataset comprised the following fields: first name, middle name, surname, date of birth, sex, address, suburb, postcode and state. This data had previously been linked to a very high standard using probabilistic linkage along with a rigorous manual review of created links, and a quality assurance program to analyse and manually review likely errors. Based on quality assurance procedures, the estimated error rate of this linkage is 0.3% [39]. Furthermore, these links have been validated through this datasets use in a large number of research projects and published research articles [1]. The links created during this original linkage allowed us to evaluate our linkage quality in comparison.

Both synthetic data and real administrative data have advantages and disadvantages comparison data sets. Synthetic data may not manage to capture all the complexity of errors that real administrative data can. Using real administrative data requires relying on the results of previous linkages as a standard by which to compare which may not be entirely accurate, whereas synthetic data gives a known, accurate standard. By using both of these datasets in our analysis, we hope to avoid both of these issues, and gain the best of both worlds.

Cleaning techniques

For each dataset, two sets of cleaned variables were computed – a minimally cleaned set and a heavily cleaned set. Information on the specific techniques used in each dataset can be found in Table 3. The generation of some variables required the creation of additional lookup tables: a nickname table, and a sex imputation table.

A nickname lookup table was developed based on similar nickname lookup tables found in linkage packages and as used by Australian linkage units. A sex imputation table was developed by examining the frequency of each given name in the data files and calculating the probability of the person being male or female. A record with a missing sex value was then given the most common gender value for this name.

Linkage strategy

The linkage strategy chosen was based on a previously published default strategy used for an evaluation of linkage software [38]. A probabilistic linkage approach was used with two blocks (Soundex of surname with first initial, and date of birth) and all possible comparison variables were computed in each block. A String similarity measure (the Jaro-Winkler string comparator [40]) was

Table 2 A comparison of the most common fields in the created synthetic data and the original data it was based on

Surname (top 5)	Synthetic	Original	Male forename (top 5)	Synthetic	Original
	Per cent	Per cent		Per cent	Per cent
Missing value	1.98		Missing value	1.99	
Smith	0.92	0.94	John	3.44	3.47
Jones	0.55	0.55	David	3.09	3.09
Brown	0.46	0.46	Michael	2.95	2.95
Williams	0.46	0.46	Peter	2.87	2.88
Taylor	0.44	0.44	Robert	2.47	2.47
Female forename (top 5)	**Synthetic**	**Original**	**Postcode (top 5)**	**Synthetic**	**Original**
	Per cent	**Per cent**		**Per cent**	**Per cent**
Missing value	1.99		Missing value	1.01	
Margaret	1.57	1.56	6210	2.84	2.84
Susan	1.35	1.34	6163	2.33	2.34
Patricia	1.22	1.22	6027	2.06	2.05
Jennifer	1.19	1.20	6155	2.02	2.02
Elizabeth	1.05	1.05	6065	2.00	1.98

Table 3 Specific data cleaning techniques used on each dataset

Synthetic data

Fields available for linkage: forename, surname, date of birth, sex, postcode

No cleaning	**Minimal cleaning**	**High cleaning**
Reformat values:	**Reformat values:**	**Reformat values:**
Not required	Not required	Not required
	Remove alt. missing values and uninformative values:	**Remove alt. missing values and uninformative values:**
	Invalid dates of birth removed	Invalid dates of birth removed
	Invalid postal code values removed	Invalid post code values removed
	Remove punctuation:	**Remove punctuation:**
	Both forename and surname fields had all punctuation and spaces removed	Both forename and surname fields had all punctuation and spaces removed
		Nickname lookup:
		Nicknames were changed to their more common variant.
		Sex Imputation
		Records with missing sex had a value imputed based on their first name.

Hospital admissions data

Fields available for linkage: forename, middle name, surname, sex, date of birth, address, suburb, postcode, state

No cleaning	**Minimal cleaning**	**High cleaning**
Reformat values:	**Reformat values:**	**Reformat values:**
Date of birth reformatted.	Date of birth reformatted	Date of birth reformatted.
	Remove alt. missing values and uninformative values:	**Remove alt. missing values and uninformative values:**
	Invalid dates of birth were removed	Invalid dates of birth were removed
	Invalid postcode values were removed ('9999' etc.)	Invalid postcode values were removed ('9999' etc.)
	Uninformative address and suburb values removed ('NO FIXED ADDRESS', 'UNKNOWN' etc.)	Uninformative address and suburb values removed ('NO FIXED ADDRESS', 'UNKNOWN' etc.)
	Birth information encoded in first name removed ('TWIN ONE OF MARTHA' etc.)	Birth information encoded in first name removed ('TWIN ONE OF MARTHA' etc.)
	Remove punctuation:	**Remove punctuation:**
	Forename, middle name surname and suburb fields had all punctuation and spaces removed	Forename, middle name surname and suburb fields had all punctuation and spaces removed
		Nickname lookup:
		Nicknames were changed to their more common variant.

used for all alphabetic variables (names, address and suburb) with exact matches being carried out on all other variables. Day, month and year of birth were all compared separately. Correct agreement and disagreement weights for probabilistic linkage [41] were calculated for each variable and used in linkage. The threshold setting was adjusted multiple times with the linkage quality computed for each adjustment, with the highest result (i.e. the largest F-measure) reported. The threshold was adjusted in both directions in increments of 0.5, until it was clear all future adjustments would continue to worsen the F-measure. This linkage strategy was based on a previously published 'default' linkage strategy [38].

Linkage methods

As probabilistic record linkage techniques provide robust matching results for data which contain inconsistencies or incomplete data, these have been used throughout the study to match both the synthetic and 'real world' data sets. Following the traditional probabilistic linkage approach, pairs of records were compared and classified as matches if the matching score is above the threshold.

To calculate the matching score reached by a pair of records, each field (for instance first name or postcode) has been compared. Scores for each individual field were computed using agreement and disagreement weights. The agreement weight expresses the likelihood that records which belong to the same person have the same value for this field. The disagreement weight expresses the likelihood that records which do not belong to the same person have the same value on this field. The sum of these individual field scores has been computed and compared to the matching threshold to determine matches or non-matches [15].

Linkage engine

BigMatch, developed by the US Bureau of Census [42] was used as the linkage engine for the analysis. BigMatch was chosen as it is fast, can handle large volumes, has a transparent linkage process based on probabilistic methods, and importantly, does not contain any automatic inbuilt data cleaning. The software had previously been evaluated and found to perform well against other linkage software packages [38].

Measuring linkage quality

There are two types of errors that can be made in record linkage. Firstly there are incorrect matches, whereby two records are designated as belonging to the same person when they should not be (a false positive). Secondly there are missed matches, whereby two records are not designated as belonging to the same person when they should be (a false negative). These two types of errors can be measures as precision (the proportion of matches found that were correct) and recall (the proportion of correct matches that were found). A linkage with a high precision will have few false positives; similarly a linkage with high recall will have few false negatives. The F-measure of a linkage is the harmonic mean between precision and recall. This gives us a single equation with which we can compare linkage quality. These measures have been recommended as suitable for record linkage [43], and have been used previously in record linkage studies [38]. The calculations for these measures can be seen below.

$$Precision = \frac{Total\ number\ of\ correct\ pairs\ found}{Total\ number\ of\ pairs\ found}$$

$$Recall = \frac{Total\ number\ of\ correct\ pairs\ found}{Total\ number\ of\ correct\ pairs}$$

$$f - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Measuring the quality of a single variable

A similar approach to the one described above can be used when measuring the quality of a single variable. A variable which nearly always has the same value for all records belonging to the same person, but nearly always has a different value than all records belonging to other people, would be much more useful in the linkage process than one which seldom had these properties. Put in another way, a variable with a high precision (here measured as the proportion of times that two variables which have the same value belong to the same person) and a high recall (the proportion of times two records matching each other had the same value of the variable in question) will be more useful than one with lower precision and recall.

As some data cleaning techniques may increase precision and lower recall, we can determine which technique will have the overall best effect on predictive accuracy by using the F-measure of these two values. Furthermore we can measure the relative improvement of a data cleaning technique by comparing its individual F-measure before and after data cleaning.

Results

The overall linkage quality results can be seen in Table 4. This represents the highest possible F-measure in each cleaning condition after testing multiple thresholds. The differences found when manipulating the level of data cleaning were very small. For both synthetic and hospital admissions data, a high level of data cleaning resulted in a decrease in linkage quality. Minimal cleaning resulted in a slight decrease in linkage quality for synthetic data, while remaining the same for hospital admissions data.

Data cleaning techniques were further investigated to determine their individual effect in improving or decreasing linkage quality. Each variable had its predictive ability determined by calculating its own precision, recall and F-measure, where two values were said to match if they were exactly the same. The percentage difference in predictive ability between the cleaned variables and the

Table 4 Overall linkage quality results

Synthetic data	
	F-measure
No cleaning	0.883
Minimal cleaning	0.882
High cleaning	0.875
Hospital admissions data	
	F-measure
No cleaning	0.993
Minimal cleaning	0.993
High cleaning	0.992

Table 5 Improvement in predictive ability of data cleaning techniques

	Hospital admissions data	Synthetic data
Remove punctuation	–[a]0.08%	+0.08%
Remove alt. missing values	+0.5%	0%
Nickname lookup	–28%	–33%
Sex Imputation	NA	–5%

[a] Negative sign (-) refers to decrease in predictive ability, positive sign (+) refers to increase in predictive ability compared to baseline.

original variables was then computed, with the average percentage change for each cleaning technique shown in Table 5. As there were no missing values for sex in the hospital admissions data, this technique was not used.

While removing missing values and uninformative values seemed to increased predictive ability, all other techniques displayed mixed or worse results. Using name variables that had nicknames and diminutive names replaced with their original names resulted in a large 30% decrease in that variable's predictive value.

A sample of the precision and recall of the variables used is shown in Table 5. For individual transformations, the amount of correct matches found typically increases with data cleaning (increased recall), while the number of incorrect matches found also increases, resulting in lower precision. In general, the decrease in precision more than offsets the increase in recall, resulting in a decreased overall result. For instance, while the Soundex of surname (Table 6) resulted in an increase in the amount of correct matches found compared to the original surname field (from 98.8% to 99.4%, an increase of 0.6%), the percentage of matches found that were correct dropped 65% from 2.53% to 0.88%. This pattern is seen for most other transformations, and appears to be the reason for the decrease in linkage quality.

Discussion

Overall, it was found that the effect of data cleaning on linkage quality was very small. If there was any effect at all, it appeared to decrease linkage quality. While some techniques led to small improvements, many others led to a large decrease in quality.

These results were not as expected. Data cleaning is assumed to improve data quality and thus to increase linkage quality. Examining the effect individual transformations had on a single variable's predictive ability allows us to explain why this occurred. While the number of correct matches that were brought together increased with data cleaning, the number of incorrect matches also increased, in most cases dramatically. By removing the variability between records we are reducing our ability to distinguish one record from another.

Data cleaning techniques typically reduce the variability between values of the field in question. By removing nicknames, a smaller variety of names will be found in the dataset. By removing differences created by punctuation, this variability will be removed. As anticipated [7] this leads to a greater number of correct matches found; however this also leads to the identification of more incorrect matches.

Strengths and limitations

Given the acceptance of data cleaning as an integral part of the linkage process, it was assumed that data cleaning would improve quality in general. The results obtained appear to contradict the conventional wisdom that data cleaning is a worthwhile procedure due to its ability to improve linkage quality.

Through the use of multiple representative datasets and the analysis of both linkage quality and individual transformations, these results seem robust. Measuring the effect of data cleaning in linkage is complex, as there are a multitude of parameters which can be altered that could affect the outcome of linkage quality. A potential

Table 6 Examples of single variable changes in predictive ability for individual cleaning techniques in hospital admission data

Hospital admissions data			
	Precision	**Recall**	**F-measure**
Percentage difference from original variable			
Given name original	0.006575	0.946085	0.013059
Given name with removed punctuation	0.006573↓[b]*0.03%*	0.947188↑*0.11%*	0.013056↓*0.02%*
Given name with nicknames removed	0.004357↓*33.7%*	0.953738↑*0.81%*	0.008675↓*33.5%*
Surname original	0.025265	0.98824	0.049271
Soundex of surname	0.008845↓*65%*	0.994926↑*0.67%*	0.017533↓*64.4%*
Address original	0.687066	0.669649	0.678246
Address with alternate missing values and uninformative values removed	0.687398↑*0.05%*	0.709426↑*5.9%*	0.698238↑*2.9%*

[b] Down arrow symbol (↓) refers to decreased percentage change, up arrow (↑) refers to increased percentage change.

concern is that some untested threshold value or other linkage parameter changes could drastically change these results. However, when analysed on their own, individual variables showed decreased predictive ability. If we accept that record linkage variables are independent (something which is an assumption of probabilistic record linkage) then it seems unlikely that any changes to linkage parameters will lead to linkage quality greater than that found in uncleaned data. On the other hand, the independence of variables used in linkage is often questionable, in which case the lower predictive ability of the individual variables is at the very least supportive of our conclusion.

The linkage strategy adopted here made heavy use of string similarity metrics. String similarity metrics may reduce the need for data cleaning, as they allow finer grained measures of similarity compared to exact matching, where variables with very slight differences will be treated as non-matches. A linkage strategy using exact matching only will have more need for data cleaning to bring correct records together, and this linkage strategy was not tested. However, the analysis of predictive ability of individual variables and their cleaned versions was carried out with exact matching only, which showed a decrease in predictive ability. This suggests data cleaning would not affect results any differently for those using an exact matching linkage strategy.

The linkages conducted simply replaced the original variables with the cleaned variables. An alternative method may be to use both the original and cleaned versions as variables in linkage. While this method violates the assumptions of independence underlying probabilistic record linkage [41], linkage variables are almost never independent, and such techniques have been implemented in some linkage packages. Further work would be required to determine the effect of using cleaned variables in conjunction with original uncleaned variables.

The f-measure was used as the sole measure of linkage quality. An underlying assumption of using this measure is that a single false positive is as equivalently undesirable as a single false negative. While this seems a sensible starting point, it should be noted that in numerous practical applications of record linkage this is not the case. For instance, if linking registry information to inform patients of their condition, it is much more important to reduce false negatives than false positives. Further analysis using additional metrics may be required to ensure these results hold using other linkage quality metrics. The key reason why cleaning failed to improve quality was the reduced variability of each field. Other data cleaning techniques not investigated here such as address standardisation increase the number of variables available for comparison and these techniques may improve quality.

Avenues for further research

From this work it is clear that data cleaning does not always lead to increased linkage quality. Without further testing on a wide variety of datasets, it is hard to draw any further conclusions about the use of data cleaning in record linkage. Repeating this research on a wide variety of datasets is important. Further research into the use of cleaned as well as uncleaned variables together in the same linkage, into the use of further cleaning technique such as name and address standardisation is required. This research suggests that there are some situations where data cleaning transformations are helpful and others where they are not – determining a way of identifying when a transformation is likely to be helpful would be an important and useful finding.

Conclusion

Data cleaning encompasses a variety of techniques which will be appropriate in specific circumstances. Care should be taken when using these techniques.

Additional file

Additional file 1: Contains the synthetic data used in this paper. This file is in comma separated, delimited format and is viewable in Microsoft Excel or any text editor. The features of this dataset are described more fully in the manuscript.

Competing interests

All researchers involved in this study are employed by an Australian university. As with all Australian universities, the publication of work in a peer reviewed journal will result in credit being received.

Authors' contributions

Initial idea for research developed by AF. Linkage and analysis conducted by SR, with input from AF, JS and JB. SR drafted the manuscript with JS, JB and AF all providing substantial contributions. All authors read and approved the final manuscript.

Acknowledgements

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network. The authors would like to thank the reviewers for their invaluable comments.

Received: 17 March 2013 Accepted: 29 May 2013
Published: 5 June 2013

References

1. Brook EL, Rosman DL, Holman CDAJ: **Public good through data linkage: measuring research outputs from the Western Australian data linkage system.** *Aust N Z J Public Health* 2008, **32:**19–23.
2. Hall SE, Holman CDAJ, Finn J, Semmens JB: **Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records.** *Int J Qual Health Care* 2005, **17:**375–381.
3. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB: **Data linkage infrastructure for cross-jurisdictional health-related research in Australia.** *BMC Health Serv Res* 2012, **12:**480.
4. Fellegi IP, Sunter AB: **A theory for record linkage.** *J Am Stat Assoc* 1969, **64:**1183–1210.

5. Pinder R, Chong N: **Record linkage for registries: current approaches and innovative applications.** http://www.naaccr.org/LinkClick.aspx?fileticket=wtyP5M23ymA%3D.
6. Gomatam S, Carter R, Ariet M, Mitchell G: **An empirical comparison of record linkage procedures.** *Stat Med* 2002, **21:**1485–1496.
7. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB: **Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage.** *J Clin Epidemiol* 2011, **64:**565–572.
8. Clark DE, Hahn DR: **Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry.** In *Proceedings of the annual symposium on computer application in medical care.* Maryland USA: American Medical Informatics Association; 1995:397.
9. Newcombe HB, Smith ME, Howe GR, Mingay J, Strugnell A, Adbatt JD: **Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers.** *Comput Biol Med* 1983, **13:**13.
10. Roos LL JRAW, Nicol JP: **He art and science of record linkage: methods that work with few identifiers.** In *Book the art and science of record linkage: methods that work with few identifiers.* Winnipeg, Canada: Departments of Business Administration and Social and Preventive Medicine University of Manitoba; 1985.
11. Roos L, Wajda A: **Record linkage strategies. Part I: Estimating information and evaluating approaches.** *Methods Inf Med* 1991, **30:**117.
12. Quantin C, Bouzelat H, Allaert F, Benhamiche A-M, Faivre J, Dusserre L: **How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure.** *Int J Med Inform* 1998, **49:**117–122.
13. Wajda A, Roos LL: **Simplifying record linkage: software and strategy.** *Comput Biol Med* 1987, **17:**239–248.
14. Gu L, Baxter R, Vickers D, Rainsford C: **Record linkage: current practice and future directions.** *CSIRO Mathematical and Information Sciences Technical Report* 2003, **3:**83.
15. Herzog TN, Scheuren FJ, Winkler WE: *Data quality and record linkage techniques.* New York: Springer; 2007.
16. Winkler WE: *Record linkage software and methods for merging administrative lists.* Statistical research division Technical Report RR01—03 US Bureau of Census; 2001.
17. Jaro MA: **Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.** *J Am Stat Assoc* 1989, **89:**414–420.
18. Gill L: *Methods for automatic record matching and linkage and their use in national statistics.* London, UK: Office for National Statistics; 2001.
19. Ferrante A, Boyd J: **Data linkage software evaluation: a first report (part I). Perth.** In *Book data linkage software evaluation: A first report (part I).* Perth: Curtin University; 2010.
20. Christen P: *Data matching.* New York: Springer; 2012.
21. Odell KM, Russell RC: **Soundex phonetic comparison system.** vol. 1261167th editionUS Patent 1261167; 1918.
22. Taft RL: **Name search techniques.** New York: Bureau of Systems Development; 1970.
23. Philips L: **Hanging on the metaphone.** *Computer Language* 1990, 7(23):39–42.
24. Day C: *Record linkage II: experience using AUTOMATCH for record linkage in NASS.* USA: US Department of Agriculture; 1996.
25. Churches T, Christen P, Lim K, Zhu JX: **Preparation of name and address data for record linkage using hidden markov models.** *BMC Med Inform Decis Mak* 2002, **2:**9.
26. **IBM Infosphere QualityStage.** http://www-01.ibm.com/software/data/infosphere/qualitystage/.
27. **Linkage Wiz data matching software.** http://www.linkagewiz.net/.
28. Borthwick A, Buechi M, Goldberg A: **Key concepts in the choicemaker 2 record matching system.** In *Procs first workshop on data cleaning, record linkage, and object consolidation, in conjunction with KDD.* Washington DC: SIGKIDD; 2003.
29. Christen P, Churches T, Hegland M: **Febrl–a parallel open source data linkage system.** In *Advances in knowledge discovery and data mining.* New York: Springer; 2004:638–647.
30. Jurczyk P, Lu JJ, Xiong L, Cragan JD, Correa A: **FRIL: A tool for comparative record linkage.** In *AMIA annual symposium proceedings.* Maryland, USA: American Medical Informatics Association; 2008:440.
31. Campbell KM, Deck D, Krupski A: **Record linkage software in the public domain: a comparison of link plus. The link king and a 'basic' deterministic algorithm.** *Health Informatics* 2008, **14:**5–15.
32. Howe G, Lindsay J: **A generalized iterative record linkage computer system for use in medical follow-up studies.** *Comput Biomed Res* 1981, **14:**327–340.
33. Yancey WE: **BigMatch: a program for extracting probable matches from a large file for record linkage.** *Computing* 2002, **01:**1–8.
34. Tuoto T, Cibella N, Fortini M, Scannapieco M, Tosco L: **RELAIS: Don't Get lost in a record linkage project.** In *Proc of the federal committee on statistical methodologies (FCSM 2007) research conference.* Arlington, VA, USA: Federal Committee on Statistical Methodologies; 2007.
35. Winkler WE (Ed): *Matching and record linkage.* New Jersey, USA: John Wiley & Sons; 1995.
36. Wilson DR: **Name standardization for genealogical record linkage.** In *Proc of the 5th Annual family history technology workshop.*. USA: Brigham Young University; 2005.
37. Pudjijono A, Christen P: **Accurate synthetic generation of realistic personal information.** In *Proceedings of the 13th pacific-asia conference on advances in knowledge discovery and data mining.* USA: Springer; 2009.
38. Ferrante A, Boyd J: **A transparent and transportable methodology for evaluating data linkage software.** *J Biomed Inform* 2012, **45:**165–172.
39. Rosman D, Garfield C, Fuller S, Stoney A, Owen T, Gawthorne G: **Measuring data and link quality in a dynamic multi-set linkage system.** In *Book measuring data and link quality in a dynamic multi-set linkage system.* WA: Data Linkage Unit, Department of Health; 2001.
40. Winkler WE: *String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage*; 1990.
41. Newcombe HB: *Handbook of record linkage: methods for health and statistical studies, administration and business.* New York: Oxford University Press; 1988.
42. Yancey WE: **BigMatch: a program for extracting probable matches from a large file for record linkage.** Maryland USA: Statistical Research Division U.S. Bureau of the Census; 2002:01.
43. Christen P, Goiser K: **Quality and complexity measures for data linkage and deduplication.** In *Quality measures for data mining. Volume 43.* Berlin: Springer; 2007:127–151. *Studies in Computational Intelligence.*

doi:10.1186/1472-6947-13-64
Cite this article as: Randall *et al.*: The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making* 2013 **13**:64.

Publication 11

Randall, S. M., Boyd, J. H., Ferrante, A. M., Brown, A. P., Semmens, J. B. (2015). **Grouping methods for ongoing record linkage**. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, performed all analyses, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

Grouping methods for ongoing record linkage

Sean M. Randall
Centre for Data Linkage
Curtin University
Perth, Australia
sean.randall@curtin.edu.au

James H. Boyd
Centre for Data Linkage
Curtin University
Perth, Australia
j.boyd@curtin.edu.au

Anna M. Ferrante
Centre for Data Linkage
Curtin University
Perth, Australia
a.ferrante@curtin.edu.au

Adrian P. Brown
Centre for Data Linkage
Curtin University
Perth, Australia
adrian.brown@curtin.edu.au

James B. Semmens
Centre for Population Health Research
Curtin University
Perth, Australia
james.semmens@curtin.edu.au

ABSTRACT

The grouping of record-pairs to determine which records belong to the same individual is an important part of the record linkage process. While a *merge* grouping approach is commonly used, other methods may be more appropriate when linking to a repository of previously linked data.

In this paper, we applied a number of grouping strategies to three large scale hospital datasets (comprising around 27 million records), each with a known truth set. These datasets were linked against a created 'repository' whose quality was varied.

Experimental results show that alternate grouping methods can yield very large benefits in linkage quality, especially when the quality of the underlying repository is high. *Best link* methods can remove between 25-90% of matching errors, depending on the characteristics of the underlying datasets.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Record linkage, grouping

This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo'15), Sydney, 10 August 2015. Copyright of this work is with the authors.

1. INTRODUCTION

Widely utilised in health research, record linkage involves identifying records which belong to the same individual within and across administrative datasets. By linking together records from hospital and emergency collections, primary care facilities, and birth, death and disease registries, researchers can construct a chronological sequence of events for a particular individual. The linkage process provides researchers with an enriched, cost effective, longitudinal research dataset for the study of entire populations.

In the absence of a unique identifier, linkage involves matching records using personal identifiers (e.g. name, address, and date of birth). As this information changes, and/or can be in error, statistical techniques are used to ensure links of the highest quality [4]. Ensuring high quality is critical in record linkage, as research outcomes can be affected. Current methods used to maintain linkage quality [15, 3] are heavily manual which is both costly and time-consuming. Identifying methods to improve quality that do not rely on manual review is of high interest [12].

Specialised linkage units often provide the infrastructure and expertise required to carry out record linkage. These units carry out linkage on an on-going basis, creating a list of all records and the person identifier to whom they belong. Incoming datasets are linked to the repository which is updated with this new information.

During the linkage process, incoming data is first cleaned to ensure consistency and reliability. The files are then matched using a defined linkage strategy, resulting in pairs of records designated as belonging to the same person. A grouping or clustering process then amalgamates these record-pairs into groups to identify the full set of records belonging to the same individual.

The traditional grouping process uses transitive closure to merge all identified record-pairs, with all connected records being assigned to the same individual. Transitive or *indirect* links are formed where records which did not form a pair relationship nonetheless are assigned to the same individual, for instance because they form record-pairs with a third record.

The merge based grouping process treats the repository as simply another set of records. However there is reason to believe that existing groups of records within the repository should rarely be merged together by incoming records - these groups have already been validated and are unlikely to be in error.

2. OBJECTIVES

We hypothesise that the use of grouping methods which reduce or remove the opportunity for groups within a repository to be joined together should result in higher linkage quality than the traditional merge based method. One such method has been suggested previously [9]; however this method (*best link* grouping) has never been evaluated against the traditional merge approach used in many operational linkage units across the world.

In this paper, we present an alternate best-link algorithm for grouping, and evaluate this algorithm against both the merge based and best link algorithms using real world datasets. We hypothesise that the appropriateness of these grouping techniques for on-going linkage will depend on the overall quality of the repository used. To test this, repositories of differing quality were used in the evaluation to allow us to determine the circumstances in which particular methods are appropriate.

3. METHODS

3.1 Grouping Methods

3.1.1 Merge Based Grouping

Merge grouping amalgamates all record pairs above the accepted threshold, with all connected records belonging to the same individual. Indirect or transitive links are formed where records which did not form a pair relationship nonetheless are marked as belonging to the same individual, for instance because they are both linked to a third record. If multiple groups in the repository are linked together in this way, these are merged. There is no limit to the length of indirect links accepted, although this can be used as a potential indicator of groups containing errors [12].

3.1.2 Best Link

In the approach presented by Kendrick [9], grouping is carried out in the order in which the records are matched. Each record from the incoming file is matched in turn against records in the repository. If the record from the incoming file matches to multiple records in the repository file, only the highest weighted match is accepted, and the record from the incoming file is added to this group. If the record does not link to any records in the repository, a new group is created, of which it is the sole member. The incoming record is then added to the repository, and subsequent records in the incoming file are able to match against this added record.

3.1.3 Weighted Best Link

Our modified grouping strategy which we will refer to as weighted best link, involves a linkage of records from the incoming file to the repository (along with a de-duplication of the incoming file) where all record pairs are created and evaluated. Once the linkage is completed, accepted record pairs are amalgamated in weight order. The pairs are examined in order from highest to lowest; a record-pair is accepted as valid provided it does not result in multiple groups from the repository merging together.

Algorithm 1 Best link

```
Input: Incoming file, Repository
for each record in Incoming File do
    link record to Repository
    if there is one pair found then
        add record to that group
    else if there are multiple pairs found then
        choose the highest pair
        add record to that group
    else if there are no pairs found then
        mark record as belonging to a new group
    add record to Repository
```

Algorithm 2 Weighted best link

```
   Input: Incoming file, Repository
1: Link Incoming file to Repository
2: Deduplicate Incoming file
3: Concatenate pairs from (1) and (2)
4: Sort output of (3) in weight descending order
5: for each pair in sorted pairs do
6:     if accepting will merge two repository groups then
7:         ignore pair
8:     else
9:         accept pair
```

Both best link methods assume that record-pairs have some ordinal attribute which identifies how likely they are to belong to the same individual. In probabilistic linkage, this is the weight attached to each record-pair [11]. For deterministic linkage (another common method of record linkage), these grouping strategies can be used by ordering rules by strictness.

Both best-link algorithms are similar, and in many situations return the same results. An example of their difference is shown in Figure 1. Using the best link approach, the first record A is matched to record Z and joins this group. The second record B matches to both A and Y. Of these, A is the highest weighted, so record B will join the same group as A and Z. In the weighted best link method, the first accepted pair is that joining the incoming records A and B. The next pair joins B and Y; A, B and Y are now linked together. The final pair linking A to Z is ignored, as this would bring together two groups from the repository.

The advantage of the modified weighted best link methods is that it will consistently produce the same results irrespective of the order of records being processed. The best link method described by Kendrick [9] will produce different grouping results if the linkage of the incoming records is executed in a different order.

3.2 Evaluation Datasets

Three large hospital admissions datasets were used in this evaluation, for which we had pre-existing and accurate information about which records belonged to the same person.

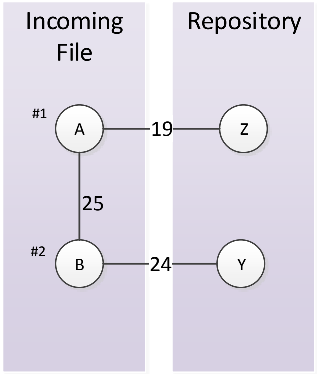


Figure 1: An example of the difference between best link algorithms. The number between records represents the weight of the record-pair comparison.

This information acted as the 'truth set' for each dataset and was used to compute differences in the performance of the three grouping algorithms. Ten years of Western Australian (WA) Hospital Admissions data, along with ten years of New South Wales (NSW) Admitted Patient Data and eight years of South Australian (SA) Hospital Admissions data were used in the evaluation. These datasets contained the typical data quality errors found in administrative data, including misspellings, name variations, missing data, changes in personal identifiers and incorrect values. Each dataset had been previously de-duplicated (by the WA Data Linkage Branch [8], the Centre for Health Record Linkage [10], and SA-NT DataLink respectively) utilising a variety of methods including exact matching, probabilistic linkage and intensive clerical review. All the linkage units employ rigorous manual reviews of created links, and a quality assurance program to analyse and review likely errors [3, 15] These links are further validated through use in a large number of research projects and published research articles [2]. Both WA and NSW have been operational for many years while in comparison SA data has only recently been linked, and has therefore been subject to less review by both clerical assessors and researchers. The data was made available as part of the Population Health Research Network Proof of Concept project [1]. A summary of the datasets is provided in Table 1.

3.3 Matching Strategy

A single matching strategy was used for all linkages in the study. This strategy utilised a probabilistic approach and was based on a previously published 'default' linkage strategy [7]. Two sets of blocks were used: Soundex of surname with first initial, and full date of birth. All variables were used in comparisons; string similarity measures were used for alphabetic variables (name, address and suburb) with exact matches used for all other variables. Agreement and disagreement weights were estimated.

3.4 Measuring Linkage Quality

Linkage quality was evaluated using saturated pairwise precision, recall and f-measure. Precision refers to the proportion of found links that were correct, and thus provides a measure of false positives. Recall is the proportion of all correct links found, and thus measures false negatives. The F-measure is the harmonic mean between precision and recall, giving a single figure from which we can compare results. These measures have been recommended for use in record linkage [5].

3.5 Repository Creation

To simulate linkage of an incoming file to a central repository, it was necessary to create *repositories* (datasets with coverage of close to the whole population). A repository for each of the original data sources was created by first randomly selecting one record per person from the hospital admissions file. This repository was 'complete' in the sense that it had coverage of the whole population being linked, and did not contain records for the same individual in more than one group.

Additional repositories of degraded quality were created by both removing records from the 'complete' repository, and by adding additional records belonging to a person already in the repository, as a separate person. Additional 'duplicate' records were specifically chosen so that differences existed in the personal identifiers between the records in the repository belonging to the same person.

Four repositories in total were created from each original dataset, differing in the number of errors they contained. These included a 'complete' repository, a repository with 1% of records missing and 1% of groups duplicated, a repository with 2.5% records missing and 2.5% groups duplicated, and a repository with 5% records missing and 5% of groups duplicated.

3.6 Evaluation Strategy

The linkage of the three datasets to their corresponding repositories was conducted separately; there was no linkage between hospital datasets.

'Incoming files' for linkage were constructed by breaking the hospital admissions records into batches containing admission records for a three month period. The batches were then linked to the repository in temporal order, to simulate on-going linkage. Records that were used to create repositories did not form part of the incoming files.

Each linkage of a batch of incoming records to the corresponding repository was grouped using three different methods - the traditional merge based method, best link and the new weighted best link approach.

Linkages were conducted using four different repositories, with three different grouping strategies, on the three state-based datasets, for a total of 36 linkage runs. The quality of each run was measured using the metrics described above.

4. RESULTS

The optimal F-measures of the overall linkage (after all batches were added) for each linkage run are shown in Figure 2. The figure displays the maximum F-measure achieved across a range of possible threshold settings.

Table 1: Dataset characteristics

Missing Values	NSW Morbidity	WA Morbidity	SA Morbidity
Surname	31.9%	<0.1%	5.3%
Given Names	33.9%	<1.0%	5.5%
Sex	<0.1%	<0.1%	<0.1%
DOB	<0.1%	<0.1%	0
Suburb	<1.0%	<1.0%	6.9%
Address	7.5%	<0.1%	8.1%
Postcode	<1.0%	<1.0%	8.5%
N	19,874,083 records	6,772,949 records	2,509,914 records

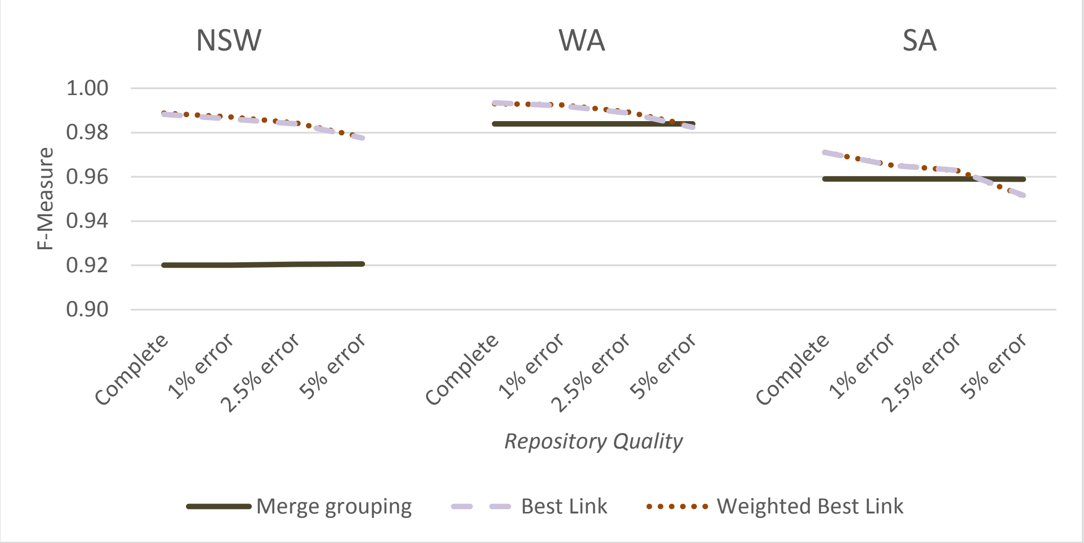


Figure 2: Results of grouping by repository quality

As can be seen, the effectiveness of merge-based grouping as compared with best link methods depended heavily on both the dataset used and the quality of the repository. For all datasets, the best link methods were superior when using a repository with an error rate of 2.5% or less. For an error rate of 5%, the most effective grouping strategy varied with the dataset.

Merge based grouping was not affected by repository quality, whereas the linkage quality of the best link methods decreased as the quality of the repository was degraded. This is unsurprising, as merge based grouping accepts all record-pairs above a certain threshold, without regard for the constitution of the repository, whereas best link methods will specifically reject certain record-pairs above the threshold based on records found in the repository.

Little difference was observed in the maximum F-measure between the two best link methods. This was a consistent finding across all datasets and all levels of repository quality.

Figure 3 shows the overall F-measure for each threshold value, for all grouping methods and for all repositories; displayed threshold are those found through probabilistic record linkage using the method of Fellegi-Sunter [6]. For higher valued thresholds, there was no difference between the merge based strategy and either of the best link strategies; however, for lower chosen thresholds the F-measures diverged, with merge based grouping scores rapidly decreasing, while best link scores improved.

As the threshold decreases, the number of false-positive pairs increase. The merge grouping method includes these false-positive pairs, resulting in lower linkage quality. Best link methods only accept these false-positives pairs if the incoming record has not already linked to a record in the repository. As this is nearly always the case, the vast majority of these false-positives are ignored, and so linkage quality remains relatively unchanged. For higher thresholds where there are fewer false-positives, there are smaller differences between these approaches.

A final notable difference is the much greater threshold range over which the F-measure for best link grouping is at a maximum.

5. DISCUSSION

The results of this study show that when optimising for linkage quality, the most appropriate grouping strategy depends on the underlying quality of the repository. If the repository is not representative of the study population or of poor quality with little confidence in the established groups, the merge based method can be considered as a possible grouping strategy. However, for better quality repositories, best link methods result in much higher linkage quality. It would be expected that most data repositories, or well-maintained

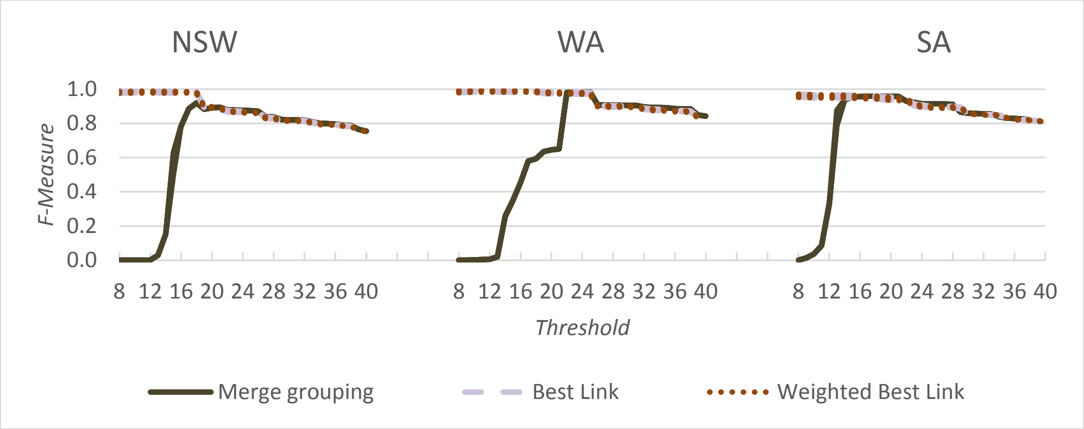


Figure 3: Results of grouping by threshold score

datasets with high population coverage, would contain only a small level of error, making best link the most appropriate grouping strategy to adopt. As the results indicate, best link methods have the added advantage of being insensitive to threshold changes. This increased tolerance reduces the likelihood of threshold estimation errors and suggests that these grouping methods could be useful in situations where determining thresholds is difficult, such as in privacy preserving linkage [13].

Our results were also highly dataset dependent, with best link methods proving superior on NSW data for all repositories. This is likely to be a reflection of the lower data quality (the NSW data has much higher rates of missing values; see Table 1).

Results showed little difference between the two best link methods. Factors other than linkage quality may be more appropriate in determining which of these methods should be used in ongoing linkage. The weighted best link method has the advantage that results are repeatable and not dependent on the order of incoming records. This means that it is possible to retrace and understand the sequence of links that were created over time without knowing the order in which records arrived. The weighted method also has the advantage that grouping decisions are made independently of matching decisions. This de-coupling of processes may be important in the design and development of linkage systems.

Given the dataset-specific nature of the results from this study, additional testing against other datasets may be required to gain a full understanding of the relationship between linkage quality, grouping strategy and population repository quality.

Our results show that the choice of grouping strategy can make a large difference to linkage quality. Within this evaluation, best link methods were able to remove between 25% (SA) to 90% (NSW) of matching errors using a high quality repository. This is an extremely large improvement in linkage accuracy, yielding far larger gains than other techniques in the literature [14, 12].

6. CONCLUSION

The effect of grouping methods on linkage quality is an understudied area of research. By adopting an appropriate grouping strategy, vast improvements in linkage quality can be achieved. The weighted best link strategy presented here shows large improvements against the merge strategy currently in operation, while providing practical benefits over the previous best link method.

Current methods of improving quality present as processing bottlenecks. Methods which improve the overall quality of linked data without impacting on performance will ultimately lead to more accurate and reliable research outcomes and increased utilisation of this resource by researchers.

7. ACKNOWLEDGMENTS

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network.

8. REFERENCES

[1] J. H. Boyd, A. M. Ferrante, C. M. O'Keefe, A. J. Bass, S. M. Randall, and J. B. Semmens. Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC health services research*, 12(1):480, 2012.

[2] E. L. Brook, D. L. Rosman, and C. J. Holman. Public good through data linkage: measuring research outputs from the western australian data linkage system. *Australian and New Zealand journal of public health*, 32(1):19–23, 2008.

[3] Centre for Health Record Linkage. Quality assurance, 2015. [Online; http://www.cherel.org.au/quality-assurance; accessed 3-June-2015].

[4] P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[5] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, pages 127–151. Springer, 2007.

[6] I. P. Fellegi and A. B. Sunter. A theory for record

linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[7] A. Ferrante and J. Boyd. A transparent and transportable methodology for evaluating data linkage software. *Journal of biomedical informatics*, 45(1):165–172, 2012.

[8] C. D. J. Holman, J. A. Bass, D. L. Rosman, M. B. Smith, J. B. Semmens, E. J. Glasson, E. L. Brook, B. Trutwein, I. L. Rouse, C. R. Watson, et al. A decade of data linkage in western australia: strategic design, applications and benefits of the wa data linkage system. *Australian Health Review*, 32(4):766–777, 2008.

[9] S. Kendrick, M. Douglas, D. Gardner, and D. Hucker. Best-link matching of scottish health data sets. *Methods of information in medicine*, 37(1):64–68, 1998.

[10] G. Lawrence, I. Dinh, L. Taylor, et al. The Centre for Health Record Linkage: a new resource for health services research and evaluation. *Health Information Management Journal*, 37(2):60, 2008.

[11] H. B. Newcombe. *Handbook of record linkage: methods for health and statistical studies, administration, and business.* Oxford University Press, Inc., 1988.

[12] S. M. Randall, J. H. Boyd, A. M. Ferrante, J. K. Bauer, and J. B. Semmens. Use of graph theory measures to identify errors in record linkage. *Computer methods and programs in biomedicine*, 115(2):55–63, 2014.

[13] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens. Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics*, 50:205–212, 2014.

[14] S. M. Randall, A. M. Ferrante, J. H. Boyd, and J. B. Semmens. The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making*, 13(1):64, 2013.

[15] D. Rosman, C. Garfield, S. Fuller, A. Stoney, T. Owen, and G. Gawthorne. Measuring data and link quality in a dynamic multi-set linkage system. In *Proceedings of the Symposium on Health Data Linkage*, 2002.

Publication 12

Randall, S. M., Boyd, J. H., Ferrante, A. M., Bauer, J. K., & Semmens, J. B. (2014). **Use of graph theory measures to identify errors in record linkage.** *Computer methods and programs in biomedicine, 115*(2), 55-63.

Contribution:

SR developed the research design and evaluation methodology for the paper, developed and selected the graph theory measures used, reviewed the literature, performed all analyses, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

Chapter 5

Methods for improving privacy protection

Research Output

Supporting Publications

13. Boyd, J. H., **Randall, S. M.**, & Ferrante, A. M. (2015). **Application of Privacy-Preserving Techniques in Operational Record Linkage Centres.** In *Medical Data Privacy Handbook* (pp. 267-287). Springer International Publishing.

Key Publications

14. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). **Privacy-preserving record linkage on large real world datasets.** *Journal of biomedical informatics, 50*, 205-212

15. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., Brown, A. P., Semmens, J. B. (2016). **Limited privacy protection and poor sensitivity: is it time to move on from the Statistical Linkage Key-581?** *Health information management journal 45*(2), 71-79.

16. **Randall, S. M.**, Brown, A. P., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2015). **Privacy preserving record linkage using homomorphic encryption.** *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

5.1. Record linkage and privacy in Australia

The issue of privacy remains a key concern of record linkage practitioners. Fundamentally, this is because record linkage generally involves the use of personal information about an individual without their direct consent. There are several reasons why consent has been considered impractical to obtain. The exclusion of individuals who refused consent would likely systematically bias any research findings [245, 246], limiting their usefulness. Many patients would likely be deceased or not contactable at the time of seeking consent [245]. The prohibitive economic cost of contacting hundreds of thousands of individuals would no doubt severely reduce the number of studies utilising linked data.

Advocates for privacy take another view. They argue that privacy protections are required to ensure patients present for treatment and are honest with their clinicians, that privacy risks are heightened in our increasingly digital age, and that there is little community awareness (and thus no mandate) for the use of private information for public benefit [247].

Privacy and Australian law

Privacy in Australia is governed by several statutes, existing at both the federal and state levels. Health specific privacy legislation also exists at the state level [67]. The overarching federal legislation, the *Privacy Act 1988,* governs the collection, storage, use and disclosure of personal information, defined as information *"about an identified individual, or an individual who is reasonably identifiable"* (Sect 6) [248].

The *Privacy Act 1988* provides a mechanism for release of data for health research without consent, where the balance of the public interest in the proposed research outweighs to a substantial degree the public interest in the protection of privacy, as determined by a human research ethics committee [249]. Similar criteria exist in state-based statutes [250].

In the absence of consent, anonymisation provides an alternate route for data release. Several decisions by the Commonwealth Privacy Commissioner and the Victorian Civil and Administrative Tribunal have asserted that anonymous health information does not meet the criteria of personal information for the purposes of privacy legislation [247]; as such, it does not have these statutory protections. The key question is as to when data can be said to be anonymised.

Case law suggests data can be considered anonymised based on whether personal identity can be reasonably ascertained from the information. This 'must not involve taking more than moderate steps' to determine identity [247 p. 75]. In this criteria, the perspective is taken from the person who will be accessing the data, rather than for instance, any member of the public (which would provide a more stringent criteria for anonymisation).

Along with anonymous and identifiable data, data can also be characterised as re-identifiable; that is, the data contains no directly identifiable information, but contains a code or key which allows the individual to be identified when connected to a specific second information source [251]. Case law suggests that re-identifiable data can be considered anonymous provided it meets the above criteria; that is, it involves more than moderate steps to determine identity [247].

In general, record linkage studies operate under the release mechanism for health research, whereby the research benefits outweigh the privacy risk. Under this mechanism, further improvements to privacy (such as the use of the 'separation principle', or the release of encrypted identifiers for linkage only, described below in Chapter 5.2) will lower the hurdle required for the research benefit to outweigh the privacy risk.

The anonymisation of data provides an alternate mechanism for data release. This may be especially useful for research which does not fall under the domain of health, for instance research into education or criminal justice. All data used in record linkage is re-identifiable, including data in a privacy preserved state; however the data can still be considered anonymised if re-identifiable [247]. It has been suggested that a simple method to ensure more than moderate steps are required to determine identity would be to create legal or career consequences (through signed agreements) for those who re-identify data – breaking the law, or putting a career in jeopardy would be considered taking more than moderate steps to undermine privacy [247].

Privacy and culture

Privacy laws exist to protect individuals; however it is not always clear how much concern there is regarding the use of private information for research without consent. It is clear that the public values research highly as evidenced by the increasing proportion of donations towards medical research [252]; the question that arises is to what extent privacy is a concern.

Numerous surveys have attempted to gauge public attitudes on this issue. The results of these are often contradictory. A recent survey by the Office of the Australian Information Commissioner found a sizable proportion of Australians were unhappy with the transfer of health information for the *treatment of their condition* without direct consent [253]. Previous surveys from the Information Commissioner found the majority did not support the use of de-identified data for medical research [254, 255]. Similarly a study commissioned by the Australian Medical Association showed that over 80% of respondents believed consent should be required before de-identified information was released for health research [256]. However, research conducted by the National Health and Medical Research Council indicated majority support for approved researchers to link information from different databases; support was higher when using de-identified data [257]. Consumer groups have shown support for record linkage [258], and the record linkage community has engaged with consumers through their inclusion within record linkage governance frameworks [259]. It has been suggested that the framing of the question may play a large role in determining responses [247]. If this is true, it could be supposed that most members of the public have reflected little about the use of collected health data, and few have strong opinions.

Public attitudes toward the importance of privacy can seem contradictory when much of modern behaviour involves the releasing of personal information. Most Australians have a social media profile; the business model of social media is the selling of personal information about individuals [1]. Similarly, many have rewards cards, which trade personal information for occasional discounts [1]. Social norms regarding the sharing of information appear to be in a state of flux, and opinions regarding personal information may not have caught up with current behaviour.

It has been noted that a culture of risk aversion exists within government [260], which has likely resulted in reduced access to government datasets for research [1]. Government agencies often choose not to release information for fear that it could be re-identified, over taking steps to mitigate this risk [1]. Mistakes within the public sector can cause significant reputational damage for the government of the day. There can also be fears over loss of control once data is released [1].

Risk aversion regarding data release is likely common across governments: surveys conducted in Canada indicate health data custodians have a much higher concern for privacy than the general public [261].

Within government too, attitudes appear to be changing. Recently, the Australian government has taken steps towards increasing the availability of public and private sector data, commencing an inquiry into the availability and use of data [1]. The report (currently in draft form) suggests numerous reforms to increase data use, including a new *Data Sharing and Release Act,* a National Data Custodian, and a number of Accredited Release Authorities across government sectors [1]. This strong push towards data release, occurring at the top levels of government, presents an enormous opportunity for the future of linked research in Australia.

5.2. Methods for ensuring privacy

The advancement of methods to further protect privacy in record linkage remains an active area of research. Efforts to ensure privacy range from governance and IT provisions, to utilising specific data flows (such as the *separation principle*) and finally in the development of privacy preserving record linkage. These efforts are discussed in detail in *Publication 13: Application of Privacy-Preserving Techniques in Operational Record Linkage Centres [222].*

The separation principle

The 'separation principle' enhances privacy by ensuring that personally identifying information is held separately to clinical information [44]. It follows the observation that access to personal identifiers without the associated clinical information dramatically reduces the risk to privacy. In some instances this privacy risk can still be large (the existence of an individual in a data collection may itself reveal sensitive information – for instance a record of an individual in a mental health collection or a cancer registry). In the best case scenario, little more information is provided than is publically available. Similarly, access to clinical information without personal identifiers does not allow easy identification of individuals.

Under the separation principle, only the data custodian has access to both personally identifying information, and clinical content data (see Figure 8). The personal identifiers are given to the linkage unit, to match against other records. The clinical information is handed to researchers as required for authorised research projects. The linkage map (the record identifier along with the newly created person identifier) is also passed to the researcher, either directly by the linkage unit, or indirectly through the data custodian. While these data flows increase the complexity of linkage, they also significantly reduce the risk to privacy. The separation principle is used throughout Australian linkage units [8, 44].

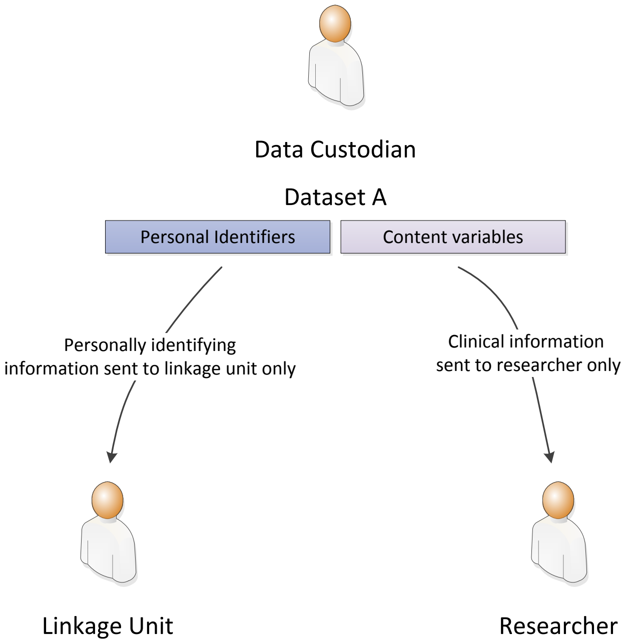

Data Custodian
Dataset A
Personal Identifiers
Content variables
Personally identifying information sent to linkage unit only
Clinical information sent to researcher only
Linkage Unit
Researcher

Figure 8: The separation principle

Privacy preserving record linkage

Privacy preserving record linkage involves carrying out record linkage on encrypted or encoded data; in this process, the linkage unit has no access to full personal identifiers, only some form of encoded information. As no personally identifying information is released by the data custodian (only encrypted identifiers to one organisation, and clinical data to another), there is a much lower privacy risk involved. These techniques can also allow record linkage to take place where there are legal requirements against the release of unencrypted information.

Privacy preserving record linkage is a popular area of current research, with a vast array of new protocols appearing, each with different aims, approaches and applicable scenarios. Comprehensive reviews of these methods exist in the literature [13].

This thesis aims to develop and evaluate practical methods for privacy preserving record linkage. To develop a method which is practical, we first need some understanding of the requirements of such a protocol *(*these requirements are discussed at greater length in *Publication 13: Application of Privacy-Preserving Techniques in Operational Record Linkage Centres [222]*).

Privacy preserving protocols can be divided into those which utilise an independent third party (such as a linkage unit) and those that do not (so called two party protocols). While two party protocols are an important development, their usefulness in Australia for health record linkage may be limited. A data custodian's role is not to conduct linkage, but to manage their specific data collections; the use of their data in linked research has no direct benefit to them. While custodians are generally accepting of the use of their datasets for linked research, it is unlikely they would be willing to dedicate the resources to conduct linkage when this does not directly add value for the custodian. Two party privacy preserving protocols are typically complex, and any such protocol would require significant time investment by data custodians. Give this background, protocols utilising an independent third party, which are far simpler for custodians, have more chance of success within a health data context.

A successful protocol will also have certain requirements with regard to efficiency, quality and security.

In practical terms, while there is no set standard for efficiency, record linkage is computationally expensive, and as dataset sizes increase, the time taken for linkage can increase dramatically. For a privacy preserving linkage method to be practical for all dataset sizes, it would need to have a runtime roughly equivalent to that of unencrypted linkage. Similarly, although privacy and security is the premise of privacy preserving protocols, there is generally no specific 'level' of privacy required; different levels of security are required in different contexts, and all else being equal, a more secure protocol is preferred over one which is less secure. Finally in terms of linkage quality, protocols which utilise approximate matching techniques along with field based weights are likely to provide much higher linkage quality than those which solely rely on exact matching on identifiers [222].

Based on these requirements, one privacy preserving method which appears to show significant promise is privacy preserving record linkage using Bloom filters [14]. This three-party protocol provides a method for approximate string comparison. Its flexibility means it could be adapted into a probabilistic record linkage framework, allowing the use of field based weights; as such it is likely to achieve high quality. As all data is hashed, no personal identifiers are released or made 'reasonably identifiable'.

Publication 14: Privacy preserving linkage on large real world datasets [200] outlines the original Bloom filter method, proposes extensions to this method to allow full probabilistic record linkage on encrypted identifiers, and tests the new method on large real-world datasets. The presented method achieves results equal to that achieved by probabilistic linkage on full unencrypted identifiers within a comparable time frame. These results suggest that privacy preserving record linkage may serve as a genuine alternative to the traditional unencrypted approach.

One linkage approach which is used in practice in Australia is the Statistical Linkage Key-581 (SLK). The SLK was developed in Australia as a method of safeguarding data privacy [15]. This 'key' is an amalgamation of components of particular personal identifiers; first and last names, date of birth and sex (see Figure 9). Records can be matched on SLK (those with exactly the same SLK are considered to be the same person). This approach thus allows record linkage to occur with somewhat obfuscated identifiers. While this approach provides some privacy protection, it is less than that provided by the Bloom filter method described above, in which all identifiers are obfuscated. As the key is an exact match on particular attributes, its linkage quality also may not be as high as that achievable using approximate matching in a probabilistic record linkage framework.

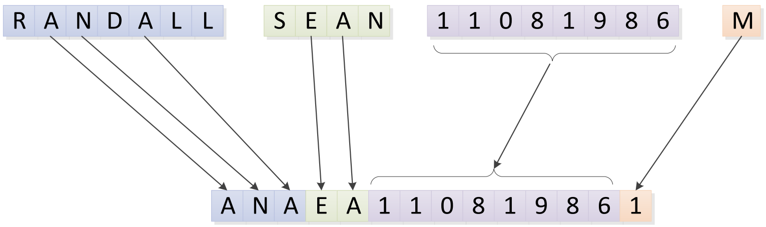


Figure 9: The Statistical Linkage Key (SLK)

The privacy safeguards and attainable linkage quality of the SLK method and the Bloom filter method are evaluated in *Publication 15: Limited privacy protection and poor sensitivity: is it time to move on from the Statistical Linkage Key-581?* [262]. The Bloom filter method is shown to achieve a higher linkage quality and provide greater privacy that the SLK method.

While the Bloom filter method provides greater privacy safeguards compared to the alternatives used in practice, it has been suggested that it may be vulnerable to frequency attacks [16]. Frequency attacks use the fact that certain identifiers are more common than others (for instance, the first name 'John') to learn information about the encrypted data. While the Bloom filter method remains superior to the currently practiced alternatives, a method which improves upon its security could have appeal and utility.

An alternate protocol with a higher level of security is presented in *Publication 16: Privacy preserving record linkage using homomorphic encryption* [263]. In this paper, a novel protocol is developed which builds on the previous Bloom filter protocol, but provides greater security guarantees; in particular, it is immune to frequency based attacks. The proposed method is shown to provide the same level of linkage quality as the previous Bloom filter method; however, a drawback is that it is slower.

The Bloom filter method presented in these publications is a clear improvement over techniques for ensuring privacy that are currently in use. Nevertheless, further research is required to ensure this method functions across a range of datasets and linkage conditions. In addition, robust methods are required to estimate appropriate linkage parameters and ensure quality; standard methods for these problems may no longer work when the data is obfuscated. However these challenges are not insurmountable. The Bloom filter method presents an important opportunity to improve the privacy safeguards of record linkage, and thereby ultimately increase access to administrative datasets.

Publication 13

Boyd, J. H., **Randall, S. M.**, & Ferrante, A. M. (2015). **Application of Privacy-Preserving Techniques in Operational Record Linkage Centres.** In *Medical Data Privacy Handbook* (pp. 267-287). Springer International Publishing.

Contribution:

SR supported the development of this paper, assisting in writing the first draft of the manuscript and contributing to the final version.

Publication 14

Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). **Privacy-preserving record linkage on large real world datasets.** *Journal of biomedical informatics, 50*, 205-212

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, developed the required software for linkage, performed all analyses, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

Publication 15

Randall, S. M., Ferrante, A. M., Boyd, J. H., Brown, A. P., Semmens, J. B. (2016). **Limited privacy protection and poor sensitivity: is it time to move on from the Statistical Linkage Key-581?** *Health information management journal 45*(2), 71-79.

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, performed all evaluations, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

Publication 16

Randall, S. M., Brown, A. P., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2015). **Privacy preserving record linkage using homomorphic encryption.** *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

Contribution:

SR developed the initial concept, formulated the research design and evaluation methodology for the paper, reviewed the literature, developed the required software, performed all evaluations, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

Privacy preserving record linkage using homomorphic encryption

Sean M. Randall
Centre for Data Linkage
Curtin University
Perth, Australia
sean.randall@curtin.edu.au

Adrian P. Brown
Centre for Data Linkage
Curtin University
Perth, Australia
adrian.brown@curtin.edu.au

Anna M. Ferrante
Centre for Data Linkage
Curtin University
Perth, Australia
a.ferrante@curtin.edu.au

James H. Boyd
Centre for Data Linkage
Curtin University
Perth, Australia
j.boyd@curtin.edu.au

James B. Semmens
Centre for Population Health Research
Curtin University
Perth, Australia
james.semmens@curtin.edu.au

ABSTRACT

The bloom filter method for privacy preserving record linkage [24] has been shown to be both efficient, and provide equivalent linkage quality to that achievable with unencoded identifiers [23]. However in some situations, the bloom filter method may be vulnerable to frequency attacks, which could potentially leak identifying information [18]. In this paper we extend the bloom filter protocol to include a homomorphic encryption step which removes the vulnerability to frequency attacks. We evaluate our method by conducting a de-duplication of emergency presentation data.

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration - Security, integrity, and protection

General Terms

Algorithms, Security

Keywords

Record linkage, privacy preserving record linkage, homomorphic encryption

1. INTRODUCTION

Record linkage is the process of identifying which person-based records from disparate data collections belong to the same individual. Throughout Australia, numerous operational record linkage units carry out this process, providing linked datasets to researchers, administrators and planners. Traditionally, linkage for research purposes has predominantly focused on the health sector, where it has had a significant impact on medical knowledge, and led to changes in health policy [5].

Administrative health data is highly sensitive, containing both medical and personal information collected about an individual during contact with health services and systems. The use of record linkage methods which implement privacy preserving techniques aims to satisfy privacy concerns regarding the release of named information, while allowing record linkage to take place.

Privacy preserving record linkage involves conducting record linkage on ‘scrambled data’, whereby records are identified as belonging to the same individual without the disclosure of personally identifying information. While these techniques provide safeguards around spontaneous recognition, they do not completely remove the privacy risk associated with large and complex datasets which are still susceptible to disclosure through unique combinations of the ‘content’ data.

Privacy preserving record linkage has recently become a popular area of research, with an array of protocols emerging. These protocols differ in their methods, maturity, practicality and suitability for large scale linkages. Comprehensive reviews of these methods exist in the literature [29].

1.1 Privacy preserving protocols - differences and requirements

Privacy preserving protocols can be divided into which utilise the data owners only (often known as two-party protocols) and those which include one or more independent third parties, who do not own data (often known as three-party protocols). Under a two-party protocol, only the organisations that hold data are involved in the linkage process. Under a three party model, data custodians provide encoded or encrypted data to an independent third party, which perform a specialised linkage of this data.

In Australia, when linking administrative data, the usefulness of two-party protocols appears limited. Two-party protocols require data custodians to take a substantial and ac-

This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo’15), Sydney, 10 August 2015.

tive part in the linkage process. However, data custodians exist to manage the quality and security of their collections and linking data is not part of their core business. While custodians are often happy for their datasets to be used for linked research, they typically do not have the resources to undertake linkage themselves, and in many cases conducting linkage does not offer them any direct benefit. At the same time, there are already a number of dedicated 'third party' linkage centres around Australia with significant expertise, and the resources to undertake record linkage [13, 1, 4].

Privacy preserving protocols also differ in the level of privacy they provide. The lowest level of privacy are provided by techniques such as the statistical linkage key (SLK) [16], which simply amalgamate personally identifying attributes (like name, date of birth and gender) into one variable in clear text. The next level of privacy techniques encodes data using hash functions so that those with access cannot learn any information directly from the encoded values; however these encoded values are vulnerable to frequency attacks, which can leak personally identifying information. A final class of privacy techniques encrypts data in such a way that it is not possible to learn any information about individuals. Such methods utilise cryptographic techniques similar to those used in modern computing. Few methods such as these exist, and those that do typically require data custodians to carry out multiple computations and communication steps [29, 7, 31].

For a privacy preserving record linkage protocol to be practical, it needs to be secure, efficient and provide high linkage quality; ideally both linkage efficiency and quality would be comparable to what can be achieved with un-encoded personal identifiers. Record linkage is computationally expensive, and while tight turnaround times are not always required for record linkage processing, slower algorithms can result in impractical processing times and unworkable solutions [10]. In addition to responsive linkage services, researcher expectations also include high quality matching to ensure they can draw the correct conclusions from their research [12].

1.2 Privacy preserving record linkage using Bloom filters

A protocol for privacy preserving linkage that appears most promising utilises Bloom filters to encode data in a way that is both efficient, and allows string similarity measures (important for ensuring high linkage quality) to be computed. The use of Bloom filters for privacy preserving record linkage was first proposed by Schnell in 2009 [24]. Since then, there have been numerous variants, extensions and evaluations of this protocol [23, 25, 19, 8, 30, 15]. The method has been shown to provide similar linkage quality to that found in probabilistic record linkage with un-encoded identifiers, and to be efficient enough for large scale linkages [23].

However recent evaluations have shown this method may be vulnerable to frequency attacks; first in its original field level form [22, 19], and then later for record level Bloom filters [18]. As such, in situations where very high levels of privacy are required, this method may not be sufficient.

1.3 Objectives of this paper

In this paper we outline an extension to the generic Bloom filter protocol, which utilises a somewhat homomorphic encryption scheme that allows us to calculate a similarity metric on fully encrypted identifiers. We implement and evaluate this method on a sample of real data sourced from hospital emergency departments.

2. PROTOCOL

2.1 Overview

Our proposed protocol is a 'four party' protocol; it utilises two independent parties to conduct linkage. One has responsibility for conducting the actual linkage (the *linker*), while the second has responsibility for decrypting the similarity score of the resulting record-pairs (the *decrypter*). In our protocol, data is first encoded into Bloom filters using the methods developed by Schnell [24]. We utilise record level Bloom filters [25] (where all fields from a record are placed within a single Bloom filter) although our method would also work with field level Bloom filters. These Bloom filters are then encrypted using the system described below, again at an individual record level. This encryption will use as input a public key supplied by the decrypting third party. This two-stage encryption process (personal identifiers encoded into Bloom filters which are then encrypted) is carried out by the data custodians. It should be noted that our protocol does not limit the number of data custodians to two; any number of data custodians can be involved in the linkage.

The encrypted data is then sent to the *linker*, who conducts the required linkage. The output of this linkage (a list of the record-pairs which have been compared along with their encrypted similarity score) is then sent to the *decrypter*, who, with possession of the private key, can decrypt the similarity score. The role of the decrypter must be separate from the linker, as giving the linker access to the private key to decrypt the encrypted similarity score would also allow them to decrypt the encrypted Bloom filters. An outline of these data movements is shown in Figure 1.

2.2 Bloom filter method

A Bloom filter is a binary vector of a set length with all values initially set to zero. Using the method outlined by Schnell [24], bigrams (overlapping sets of two letters) of personal identifiers are hashed, with their modulus taken with respect to the length of the Bloom filter. The corresponding position in the Bloom filter is then set to 1. There are several variations to this method; in our implementation all personally identifying fields (i.e. first name, surname, date of birth, sex, and address) are placed within a single large Bloom filter.

Bloom filters can be compared using typical set similarity comparisons. In this implementation we focus on the dice coefficient metric, outlined in section 2.6.

2.3 Homomorphic encryption

A homomorphic encryption scheme allows computations to be carried out on encrypted data producing encrypted results; when this encrypted data is finally decrypted, the decrypted results match the results of those same operations performed on an unencrypted version of the data. While

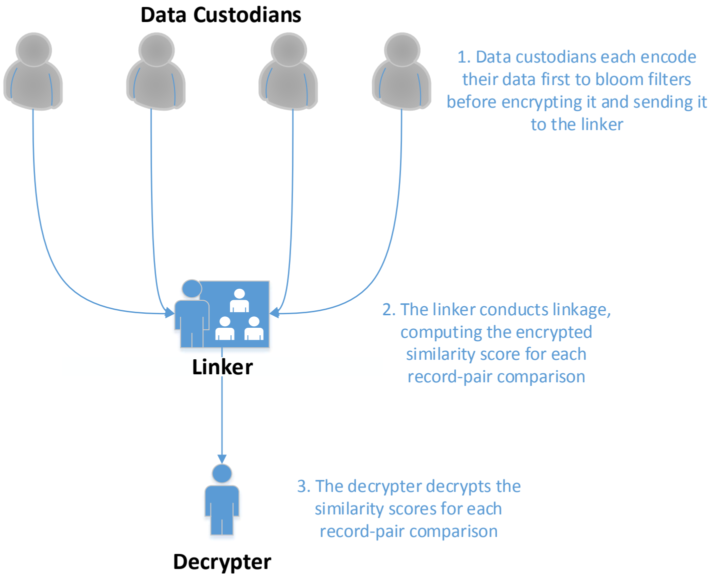

Data Custodians
1. Data custodians each encode their data first to bloom filters before encrypting it and sending it to the linker
Linker
2. The linker conducts linkage, computing the encrypted similarity score for each record-pair comparison
Decrypter
3. The decrypter decrypts the similarity scores for each record-pair comparison

Figure 1: Data movements for the proposed protocol

homomorphic encryption protocols have existed for many years, protocols prior to 2000 only supported simple operations of either addition or multiplication. In 2009, Gentry developed the first fully homomorphic encryption system which allowed arbitrary calculations [11], and since then a large number of advances in this area have been made. However fully homomorphic systems are still too slow to be practical for most purposes [20].

Somewhat homomorphic encryption schemes only support a limited number of operations on encrypted data; however they are much faster and thus far more practical. In this paper we utilise a somewhat homomorphic encryption scheme developed by Lauter, Naehrig and Vaikuntanathan [20], along with a packing method for encrypting data developed by Yasuda [32] which allows us to compute similarity measures.

2.4 Encryption method

This scheme of Lauter, Naehrig and Vaikuntanathan [20] bases its security on the *ring learning with errors* problem. In colloquial terms, this problem is based on the difficulty of distinguishing a true signal (in this case, the secret) from noisy data. The problem, while relatively recent, is believed to be exponentially hard [20], and forms the basis for numerous modern cryptosystems [2, 21].

The scheme used in this paper allows an arbitrary number of additions of encrypted values, along with a set number of multiplications.

The system utilises several parameters. These include;

- The dimension n, which is a multiple of 2, and the corresponding cyclotomic polynomial $f(x) = x^n + 1$.

- The modulus q, a prime. Together, q, n and $f(x)$ define the rings $R := \mathbb{Z}[x]/f(x)$ and $R_q := R/qR = \mathbb{Z}_q[x]/f(x)$.

- The standard deviation σ of a discrete Gaussian error distribution χ.

- An integer $t < q$, which defines the message space.

Description of the algorithms key generation, encryption and decryption are given below. These are taken verbatim from Yasuda et al [32].

Key Generation We choose an element $R \ni s \leftarrow \chi$ and sample a random element $a_1 \in R_q$ along with an error $R \ni e \leftarrow \chi$. We define the public key pk as (a_0, a_1), where $a_0 := -(a_1 \cdot s + t \cdot e)$, and we define the secret key sk as s.

Encryption For a plaintext message $m \in R_t$, with public key (a_0, a_1), the encryption samples $R \ni u, f, g \leftarrow \chi$ and computes $Enc(m, pk) = (c_0, c_1) = (a_0 u + tg + m, a_1 u + tf) \in (R_q)^2$, where $m \in R_t$ is considered an element of R_q.

Decryption For a ciphertext $ct = (c_0, ..., c_\xi) \in (R_q)^{\xi+1}$ (homomorphic multiplication will increase ciphertext size), with private key s, decryption is computed by $Dec(sk, ct) = [\widetilde{m}]_q mod\, t \in R_t$ where $\widetilde{m} = \sum_{i=0}^{\xi} c_i s^i \in R_q$.

2.5 Packing method

The homomorphic encryption scheme described above will allow us to encrypt individual numbers, and perform operations on these encrypted numbers. It is possible then to use the scheme to compute the dice coefficient of two Bloom filters, by first encrypting each element in the two Bloom filters individually, multiplying the elements of each position together, and summing these results. However such a scheme would be extremely slow, requiring a large number of encryptions and computations for every comparison.

Packing methods provide an alternative, allowing a vector of values to be encrypted in a single operation. Operations can then be homomorphically computed on this vector. In this work we utilise a packing method developed by Yasuda [32]. This method allows us to encrypt an entire Bloom filter (essentially a binary vector) at once, and compute its inner product using a single multiplication operation.

For a Bloom filter A of length n with elements $A_0, \dots, A_{n-1}$ we define two packed ciphertexts.

$$ForwardPack(A) = \sum_{i=0}^{n-1} A_i x^i$$

$$BackwardPack(A) = -\sum_{i=0}^{n-1} A_i x^{n-i}$$

where Σ refers to the regular summation operator. Both of these polynomials are then encrypted as described in 2.4. Each Bloom filter is both forward and backward packed; that is, there are two encrypted values for each Bloom filter.

We can compute the inner product of two Bloom filters by multiplying one Bloom filter's forward packing by the others backward packing, as shown below.

$$ForwardPack(A) \times BackwardPack(B)$$

$$= (\sum_{i=0}^{n-1} A_i x^i) \times (-\sum_{i=0}^{n-1} B_i x^{n-i})$$

$$= \dots - (\sum_{i=0}^{n-1} A_i B_i x^n) + \dots$$

$$= \dots + A \cdot B + \dots$$

in R_t, since $x^n = -1$ with all other terms non-constant. Thus after a multiplication, upon decryption, the value of the constant term in the resulting polynomial will be our inner product.

2.6 Computing similarity measures

The most common metric used in Bloom filter similarity calculations is the dice coefficient, typically expressed as

$$Dice\ Coefficent_{A,B} = \frac{2h}{a+b}$$

where h refers to the number of positions in both bloom filters set to 1, and a and b refer to the number of positions set to 1 in bloom filters A and B respectively.

This equation can be re-written as

$$Dice\ Coefficent_{A,B} = \frac{2A \cdot B}{A \cdot A + B \cdot B}$$

where $\cdot$ refers to the inner product operation. This allows us to compute the dice coefficient using the packing method described above.

The cryptosystem employed does not allow integer division; instead, we calculate the encrypted values of the numerator and denominator separately. Both of these values are provided (encrypted) to the *decrypter* for each record pair. Once decrypted, the *decrypter* can calculate the dice coefficient from these two provided values.

2.7 Related work

Our protocol aims to allow linkage to be conducted with only the minimum participation of data custodians, and to a level of security where frequency based information is not available to the independent third parties.

There have been a number of related works published in the literature. A range of secure set intersection protocols have been proposed [26, 27, 17], many of which adopt homomorphic encryption methods to ensure security. While these methods have strong security equivalent to our protocol, they operate without the use of an independent third party, and instead require multiple communication steps from data custodians.

The closest protocol to the one described in this paper is by Kantaricioglu et al. [14], who provides a method for privacy-preserving joins utilising homomorphic encryption and two independent third parties. Similar to our work, in this protocol data custodians are only required to encrypt and transfer their data, taking no further part in the protocol. A uniquely identifying key is used to determine whether two records should be joined. A homomorphic subtraction operation is then performed when comparing individual records; where this subtraction (when decrypted) equals to 0, the two records have the same unique identifier, and so are joined.

The main difference between our method and Kantaricioglu's is that ours is aimed at the problem of record linkage, where we do not have keys which uniquely identify individuals across distinct datasets. Our proposed method tolerates the full range of 'noisy' data, utilising approximately matching techniques to handle missing values, misspellings, incorrect values and changing values over time. Previous evaluations of the approximate matching method used in our protocol have shown it to perform as well as probabilstic linkage on un-encoded identifying information [23].

3. EVALUATION

3.1 Evaluation details

We evaluated this system by performing a deduplication of 275,626 event records (one years' worth) from an emergency presentation data collection. First name, surname, date of birth, sex, address and postcode fields were used in linkage. These fields were mapped into a single 512 bit bloom filter, using weighting methods developed by Durham et al [9]. A standard blocking method was used to enable timely linkage; the date of birth field was used as the sole block.

Bloom filters were then encrypted using the encryption scheme described above. Our system utilised the parameters $n = 1024$, $\sigma = 8$, $t = 512$, and q, a 54 bit prime. These parameters were chosen to be the most efficient possible, while both ensuring correctness of results, and a security level equivalent to 128 bits; the detail of determining ac-

Table 1: Results from de-duplication of emergency presentation data

Linkage Type	Precision	Recall	F-Measure
Linkage on un-encoded identifiers	0.985	0.978	0.981
Linkage with unencrypted bloom filters	0.985	0.977	0.981
Linkage with encrypted bloom filters	0.985	0.977	0.981

curate and secure parameters is described in Lauter et al [20].

Our linkage quality results were evaluated using precision and recall measures, as recommended in the record linkage literature [6]. Efficiency and privacy were also evaluated with reference to measures described within the privacy preserving literature [28]. The emergency presentation dataset had been previously independently linked by a data linkage unit with their results made available to us. The results were used as the 'truth set' with which we compared our results.

Encryption, linkage and decryption were performed on a 64-bit Windows Server virtual machine with an Intel Xeon E5-2609 CPU at 2.4GHz, with 32GB of memory. Our implementation utilised a single core.

3.2 Results

The results for the linkage of emergency presentation data using encrypted Bloom filters, unencrypted Bloom filters, and un-encoded personal identifiers are shown in Table 1. As expected, there was no difference in quality between encrypted Bloom filters and unencrypted Bloom filters. The Bloom filter methods result in linkage quality equal to that achieved by linkage with un-encoded identifiers.

The encrypted Bloom filter linkage took slightly over 12 hours to complete, while the encryption step took 4 hours and 20 minutes, and the decryption of the answer file took almost 17 hours. A total of 1,164,305 record comparisons were performed.

In terms of individual operations, a single inner product calculation took, on average, 31 milliseconds, while encryption of a single record took 58 milliseconds, and decryption of a single record-pair took 52 milliseconds.

Our implementation was significantly slower than the more optimised implementation reported on by Yasuda et al [32]. Using equivalent parameters, our inner product calculation (i.e. our linkage) was 27 times slower, while our encryption and decryption of data was 23 and 14 times slower, respectively. While their CPU was slightly faster (Intel Xeon X3480 at 3.07GHz), the majority of this difference appears to be due to code optimisations.

In terms of privacy, using the privacy metrics of Vatsalan [28], our protocol on its own has a degree of privacy of 0.0 (absolute privacy), as all records have completely different ciphertext values. However our protocol is not complete; for efficiency, it requires a blocking component to be used in conjunction which itself may decrease privacy.

4. DISCUSSION

As expected, the linkage quality achieved through our protocol was the same as that achieved using the regular Bloom filter method, and the same as that achieved through probabilistic linkage. The advantage of the presented methodology is a far higher level of security over the Bloom filter method. This method provides a level of security equivalent to that provided by regular encryption algorithms, and removes the possibility of frequency attacks; the same plaintext value can encrypt to a very large number of ciphertext values.

By building upon the Bloom filter methods previously published, our methodology can be expected to achieve the same level of linkage quality as other Bloom filter methods. It can also leverage off the significant work already conducted to improve and refine the Bloom filter methodology, such as Durham's weighting method (used in this paper) [9].

A key limitation to our proposed method is speed. As currently implemented, our method is only suitable for small linkages. However, our naive implementation is approximately 14 to 27 times slower than the more optimised version developed by Yasuda [32]. By optimising the code used in our implementation, our method would be suitable for larger dataset sizes. Additional performance improvements could be made by using distributed computing techniques. Given the high security level of our encryption method, it may also be feasible to utilise public cloud computing resources to perform our inner product calculations, which would provide substantial potential for scalability. The blocking method used (comparing only records with the same date of birth) is relatively strict, and similarly strict blocks may be a requirement to ensure the efficiency of this method.

5. CONCLUSIONS

As far as we are aware, this is the first record linkage protocol which provides a demonstrably high level of security, without requiring numerous communication steps by data custodians. Future developments will focus on improving performance to a comparable level with that achieved by Yasuda et al [32].

This paper presents a protocol for record comparison, and does not provide any recommendations for private blocking systems. However, a private blocking scheme is necessary for a complete private linkage system. Future work will explore the use of more secure blocking methods.

Our protocol provides protection against attacks by the third or fourth party; however it does not protect against collusion by these two parties. Should these parties collude, the security of our system reduces to that of the regular privacy preserving linkage using Bloom filters (which has been evaluated previously [18]).

6. ACKNOWLEDGMENTS

This project is supported by the Australian Government National Collaborative Research Infrastructure Strategy's Population Health Research Network.

The authors would also like to thank Wenjie Lu, whose publically available code formed an initial reference point for our implementation of this cryptosystem.

7. REFERENCES

[1] J. H. Boyd, A. M. Ferrante, C. M. O'Keefe, A. J. Bass, S. M. Randall, and J. B. Semmens. Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC health services research*, 12(1):480, 2012.

[2] Z. Brakerski, C. Gentry, and V. Vaikuntanathan. (leveled) fully homomorphic encryption without bootstrapping. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 309–325. ACM, 2012.

[3] Z. Brakerski and V. Vaikuntanathan. *Fully homomorphic encryption from ring-LWE and security for key dependent messages*, pages 505–524. Springer, 2011.

[4] E. Brook, D. Rosman, C. Holman, and B. Trutwein. Summary report: research outputs project, WA data linkage unit (1995–2003). Perth: WA data linkage unit, 2005.

[5] E. L. Brook, D. L. Rosman, and C. D. J. Holman. Public good through data linkage: measuring research outputs from the western australian data linkage system. *Australian and New Zealand Journal of Public Health*, 32(1):19–23, 2008.

[6] P. Christen and K. Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, pages 127–151. Springer, 2007.

[7] W. Du and M. J. Atallah. Secure multi-party computation problems and their applications: a review and open problems. In *Proceedings of the 2001 workshop on New security paradigms*, pages 13–22. ACM, 2001.

[8] E. Durham, Y. Xue, M. Kantarcioglu, and B. Malin. Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. *Information Fusion*, 13(4):245–259, 2012.

[9] E. A. Durham. *A framework for accurate, efficient private record linkage*. Thesis, 2012.

[10] A. Ferrante and J. Boyd. A transparent and transportable methodology for evaluating data linkage software. *Journal of Biomedical Informatics*, 45(1):165–172, 2012.

[11] C. Gentry. *A fully homomorphic encryption scheme*. Thesis, 2009.

[12] K. Harron, A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein. Evaluating bias due to data linkage error in electronic healthcare records. *BMC medical research methodology*, 14(1):36, 2014.

[13] K. A. Irvine and L. K. Taylor. The centre for health record linkage: fostering population health research in NSW. *New South Wales public health bulletin*, 22(2):17–18, 2011.

[14] M. Kantarcioglu, A. Inan, W. Jiang, and B. Malin. Formal anonymity models for efficient privacy-preserving joins. *Data & Knowledge Engineering*, 68(11):1206–1223, 2009.

[15] A. Karakasidis and V. S. Verykios. Secure blocking+ secure matching= secure record linkage. *JCSE*, 5(3):223–235, 2011.

[16] R. Karmel. *Data linkage protocols using a statistical linkage key*. Australian Institute of Health and Welfare, 2005.

[17] L. Kissner and D. Song. Privacy-preserving set operations. In *Advances in Cryptology–CRYPTO 2005*, pages 241–257. Springer, 2005.

[18] M. Kroll and S. Steinmetzer. Automated cryptanalysis of bloom filter encryptions of health records. *arXiv preprint arXiv:1410.6739*, 2014.

[19] M. Kuzu, M. Kantarcioglu, E. Durham, and B. Malin. A constraint satisfaction cryptanalysis of bloom filters in private record linkage. In *Privacy Enhancing Technologies*, pages 226–245. Springer, 2011.

[20] K. Lauter, M. Naehrig, and V. Vaikuntanathan. Can homomorphic encryption be practical? In *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*, pages 113–124. ACM, 2011.

[21] V. Lyubashevsky, C. Peikert, and O. Regev. On ideal lattices and learning with errors over rings. *Journal of the ACM (JACM)*, 60(6):43, 2013.

[22] F. Niedermeyer, S. Steinmetzer, M. Kroll, and R. Schnell. Cryptanalysis of basic bloom filters used for privacy preserving record linkage. *Journal of Privacy and Confidentiality*, 6(2):3, 2014.

[23] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens. Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics*, 50:205–212, 2014.

[24] R. Schnell, T. Bachteler, and J. Reiher. Privacy-preserving record linkage using bloom filters. *BMC Medical Informatics and Decision Making*, 9(41), 2009.

[25] R. Schnell, T. Bachteler, and J. Reiher. A novel error-tolerant anonymous linking code. Report, Working Paper Series No. WP-GRLC-2011-02. Nürnberg, Germany: German Record Linkage Center, 2011.

[26] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 639–644. ACM, 2002.

[27] J. Vaidya and C. Clifton. Secure set intersection cardinality with application to association rule mining. *Journal of Computer Security*, 13(4):593–622, 2005.

[28] D. Vatsalan, P. Christen, C. M. O'Keefe, and V. S. Verykios. An evaluation framework for privacy-preserving record linkage. *Journal of Privacy and Confidentiality*, 6(1):3, 2014.

[29] D. Vatsalan, P. Christen, and V. S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

[30] D. Vatsalan, P. Christen, and V. S. Verykios. An

efficient two-party protocol for approximate matching in private record linkage. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pages 125–136. Australian Computer Society, Inc., 2014.

[31] M. Yakout, M. J. Atallah, and A. Elmagarmid. Efficient private record linkage. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 1283–1286. IEEE, 2009.

[32] M. Yasuda, T. Shimoyama, J. Kogure, K. Yokoyama, and T. Koshiba. *Practical packing method in somewhat homomorphic encryption*, pages 34–50. Springer, 2014.

Chapter 6

Conclusion

Administrative data is an underutilised resource with the potential to dramatically increase our understanding of the nature of health and disease. Not only does it provides fine-grained detail at an individual level, but it is also collected for entire populations. The worth of this data, and the need to harness it, is beginning to be recognised in Australia and around the world. Australia has made significant investment in record linkage infrastructure since 2009, providing the country with state-wide and national linkage units to carry out record linkage and provide access to linked data for research purposes [3]. The development of this infrastructure, including the first cross-jurisdictional record linkage project in Australia, was the focus of Chapter 1.

With the expansion of linkage infrastructure in Australia, there is vast potential for linked research. Studies of disease prevalence and incidence, disease survival, risk factors for health conditions, and the effectiveness of treatments and policy changes are all possible. Two unique examples of linked research were provided in Chapter 2.

To enable linked research, there is a need to ensure both a high level of linkage quality, as well as strong safeguards to privacy. Linkage quality and privacy protection have been two central issues since the beginnings of record linkage. Accurate linkage results are vital to ensure the validity of any findings derived from the analysis of linked data. As demand for linked data increases, and the size of databases increases, manual methods for ensuring quality are no longer feasible, and alternate, automated methods are required. In today's digital age where huge quantities of data are recorded and can be made public immediately, issues of privacy are also paramount. Despite the willingness of individuals to publically release large amounts of personal information on social media, the public remains cautious regarding the use of government collected information (the so called 'privacy paradox' [1]).

This thesis presents several enhancements to current practices, along with new methods for record linkage that improve both linkage quality and privacy protection.

The improvements in linkage quality are incremental in nature, but substantial; this thesis focuses on several aspects of the record linkage process, presenting evaluations of previously existing or novel approaches. The evaluation of data cleaning methods for record linkage (*Publication 12*) has shown that this common process may be of limited use, and can in some circumstances decrease quality. This can occur because changes to identifiers typically reduce their variability, increasing the likelihood that two records belonging to different people have the same value and thus receive a higher total score. This result was not expected, and had not been previously reported. It has implications for all record linkage units currently using data cleaning methods.

The best match grouping algorithm (*Publication 13*) is a significant advancement on current methods for grouping for ongoing linkage, resulting in superior quality that is robust against poor parameterisation and lower quality data. This algorithm should be highly recommended for use whenever a linkage unit is utilising a repository model of linkage (where data is added over time to an enduring collection of links). The use of repository models is already common, and its use is likely to increase as demand for linked data increases.

This thesis includes a novel use of graph theory metrics to detect errors in record linkage (*Publication 14*). This method of error detection is highly accurate (i.e. high specificity; there are few false positives) although it is not particularly sensitive (i.e. there are many missed errors). The development of an automated method to detect likely errors (even if only a portion of the actual errors) is a large advance on current approaches, which use largely manual techniques which are far less accurate. This paper provides a first step to a method that can be refined further. An important next step is to use these methods not only to detect groups of records containing errors, but also to correct these errors. As this method does not require visible personal identifiers to identify incorrect groups, it also has the potential to be used in a privacy preserving record linkage context.

The thesis aimed to evaluate current practice and develop new methods to improve linkage quality. While the developments in this thesis are substantial, they are neither comprehensive nor complete; there are numerous avenues for further research. As discussed in Chapter 2, perhaps the clearest direction for further research is in evaluation of modern machine learning methods against the widely used, but now nearly 50 years old, probabilistic record linkage method. Further detailed evaluation of many methods within the literature is likely to be the easiest step towards improving knowledge in this area.

The reason for focusing on improving linkage quality is to ensure the outcomes of research studies using this data are valid. However research into the effect of linkage error on research outcomes is limited. This limited focus may be the result of the separation of personal identifiers from clinical information – any research on this topic will likely require manipulations of both sets of data. Further work here is required, especially to extend previous work to include linkages which involve multiple records per person.

The second aim of this thesis was the development and evaluation of methods for record linkage which do not require personal identifiers, but can instead occur using encrypted identifiers. Adopting a technique in the literature [14], privacy preserving record linkage of large scale, real world data was shown to be possible (*Publication 16*). This method achieved linkage quality as high as can be achieved with unencrypted identifiers, in a similar time frame. A comparison of this method to a similar technique currently in use (the Statistical Linkage Key) showed the new method to achieve superior linkage quality and offer greater privacy protection (*Publication 17*). An alternate protocol was presented which may further reduce risks associated with potential frequency attacks by combining the method with modern cryptographic techniques (*Publication 18*).

The developments in privacy preserving linkage contained in this thesis are considerable. For the first time, it has been shown that privacy preserving linkage can be achieved without significant degradation in linkage quality. This development has the potential to transform record linkage practice; it would dramatically reduce the risk to privacy of conducting record linkage, thus potentially increasing data access and ultimately, research. Further developments are still required to manage some of the practicalities of privacy preserving record linkage. This includes methods for validating incoming data (how to ensure received data is as expected when all fields are encrypted) and validating the results of linkage (how to 'sanity check' the output of linkage when all fields are encrypted). Methods for ensuring appropriate parameterisation are also required. These developments are unlikely to pose serious challenges to the viability of this privacy preserving method. The development of practical privacy preserving linkage methods appears particularly timely given the increased focus of government policy on utilising available data for research and policy [1].

An examination of the literature (*Chapter 2)* revealed two key issues relating to research into record linkage methodology; a lack of evaluation of potential methodological advancements, and a divide or 'disconnect' between researchers and practitioners. Techniques in use by practitioners have often received little evaluation in the literature; their efficacy is unknown. Conversely, the literature is littered with protocols and methods that have been developed but which have not been rigorously evaluated, and thus remain poorly understood and underutilised by practitioners. The robust evaluation of methods and protocols is one aspect often missing from the record linkage literature.

There are likely numerous reasons for this lack of evaluation and disconnect between researchers and practitioners. There are a lack of incentives for researchers to conduct evaluations, and real administrative data with 'answers' required for evaluation are difficult to source. Researchers may be unaware of the importance of evaluations for their work to receive adoption. Practitioners, typically existing in government, are physically separated from researchers, found in academia. Governments are typically risk averse [1], and so are not willing to invest in new technologies without significant confidence in their improvements. The communication between government and researchers may be poor.

Recognising these concerns, this thesis has attempted to bridge that divide. The work in this thesis has heavily focused both on performing rigorous evaluations of all investigated methods and techniques, as well as investigating techniques used in the field by practitioners. Data cleaning is widely used in practice, but had not previously received evaluation within the literature. Evaluations were undertaken of two algorithms adapted from the literature: the best match grouping algorithm, and the Bloom filter method for privacy preserving record linkage. Both appear to be a great improvement on previous approaches. These detailed evaluations will hopefully help influence practitioners' decisions.

All papers found in this thesis have utilised comprehensive evaluations using large real-world datasets, for which there are known answers; this is one of the key strengths of our work. Such datasets are not always available to researchers, which has likely been a limiting factor in the number of evaluations found within the literature. Greater collaboration between practitioners and researchers should improve this deficit.

The record linkage landscape, both within Australia and abroad, is changing as the importance of linked administrative data continues to be recognised. While use of linked data by researchers in Australia has been steadily growing [264], it is the shift in policy focus from the top of government towards opening access to data, as evidenced in the Productivity Commission's report into Data Availability and Use [1], that could result in significant improvements to data access and the use of linked data. Other initiatives in Australia include a push to link data from clinical domains (i.e. from patient administration systems) into national repositories, such as described in the Medical Research Future Fund's Innovation Strategy [265]. Outside of Australia, there are other initiatives such as the use of record linkage as a partial replacement for the national census, under investigation in the UK and New Zealand [266, 267]. The importance of research using linked data appears likely to increase in the coming years; as such, developments such as the ones found within this thesis may prove particularly relevant.

This thesis has aimed to improve the quality and privacy of record linkage. This is only a useful development insomuch as it serves to improve access to and quality of linked data, and therefore ultimately provide a greater number of research results, with greater confidence in their validity. While record linkage methodology is a somewhat esoteric and technical field, the research which it enables continues to improve lives through changes in health policy and clinical practice [54]. Improvements in data privacy and quality will only further this.

References

1. Australian Government Productivity Commission, *Data Availability and Use: Draft Report*. 2016.
2. Brook, E.L., D.L. Rosman, and C.D.A.J. Holman, *Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System.* Australian and New Zealand Journal of Public Health, 2008 **32**(1): p. 19-23.
3. Population Health Research Network. *PHRN - About Us*. 2016 accessed 05/09/2016]; Available from: http://www.phrn.org.au/about-us/.
4. Pedersen, C.B., *The Danish civil registration system.* Scandinavian Journal of Public Health, 2011. **39**(7 suppl): p. 22-25.
5. Holman, C.D.A.J., et al., *A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system.* Australian Health Review, 2008. **32**(4): p. 766-777.
6. O'Halloran, E., et al., *The impact of non-severe burn injury on cardiac function and long-term cardiovascular pathology.* Scientific reports, 2016. **6**.
7. Zingmond, D.S., et al., *Linking hospital discharge and death records—accuracy and sources of bias.* Journal of clinical epidemiology, 2004. **57**(1): p. 21-29.
8. Lawrence, G., I. Dinh, and L. Taylor, *The Centre for Health Record Linkage: a new resource for health services research and evaluation.* Health Information Management Journal, 2008. **37**(2): p. 60-62.
9. Rosman, D., et al. *Measuring data and link quality in a dynamic multi-set linkage system.* in *Symposium on Health Data Linkage* 2002. Sydney (NSW).
10. Gill, L., *Methods for automatic record matching and linkage and their use in national statistics*, in *National Statistics Methodological Series No. 25.* 2001, Office for National Statistics.
11. Gostin, L.O., *Health information privacy.* Cornell L. Rev., 1995. **80**: p. 451-1756.
12. Hetzel, D., *Data linkage research—can we reap benefits for society without compromising public confidence?* Australian Health Consumer, 2006. **2**: p. 27-28.
13. Vatsalan, D., P. Christen, and V.S. Verykios, *A taxonomy of privacy-preserving record linkage techniques.* Information Systems, 2013. **38**(6): p. 946-969.
14. Schnell, R., T. Bachteler, and J. Reiher, *Privacy-preserving record linkage using Bloom filters.* BMC Medical Informatics and Decision Making, 2009. **9**(41).
15. Karmel, R., *Data linkage protocols using a statistical linkage key.* 2005: Australian Institute of Health and Welfare.
16. Kuzu, M., et al. *A constraint satisfaction cryptanalysis of Bloom filters in private record linkage.* in *Privacy Enhancing Technologies.* 2011. Springer.
17. Declich, S., Carter, A. O., *Public health surveillance: historical origins, methods and evaluation.* Bulletin of the World Health Organisation, 1994. **72**(2): p. 285-304.
18. Newcombe, H.B., et al., *Automatic Linkage of Vital Records.*
19. Machado, C.J. and K. Hill, *Probabilistic record linkage and an automated procedure to minimize the undecided-matched pair problem.* Cadernos de Saúde Pública, 2004. **20**(4): p. 915-925.
20. Acheson, E. and J. Evans, *The Oxford record linkage study: a review of the method with some preliminary results.* Proceedings of the Royal Society of Medicine, 1964. **57**(4): p. 269.
21. Christen, P., et al., *Parallel computing techniques for high-performance probabilistic record linkage.* Data Mining Group, Australian National University, Epidemiology and Surveillance Branch, Project web page: http://datamining. anu. edu. au/linkage. html, 2002: p. 1-11.
22. Christen, P., *A survey of indexing techniques for scalable record linkage and deduplication.* Knowledge and Data Engineering, IEEE Transactions on, 2012. **24**(9): p. 1537-1555.
23. Newcombe, H.B. and J.M. Kennedy, *Record linkage: making maximum use of the discriminating power of identifying information.* Communications of the ACM, 1962. **5**(11): p. 563-566.

24. Fellegi, I.P. and A.B. Sunter, *A theory for record linkage.* Journal of the American Statistical Association, 1969. **64**(328): p. 1183-1210.
25. Gomatam, S., et al., *An empirical comparison of record linkage procedures.* Statistics in medicine, 2002. **21**(10): p. 1485-1496.
26. Acheson, E., *Oxford Record Linkage Study A Central File of Morbidity and Mortality Records for a Pilot Population.* British journal of preventive & social medicine, 1964. **18**(1): p. 8-13.
27. Fedrick, J., *Epilepsy and pregnancy: a report from the Oxford Record Linkage Study.* British medical journal, 1973. **2**(5864): p. 442.
28. Acheson, R.M. and A.S. Fairbairn, *Record linkage in studies of cerebrovascular disease in Oxford, England.* Stroke, 1971. **2**(1): p. 48-57.
29. Fedrick, J., *Sudden unexpected death in infants in the Oxford record linkage area. An analysis with respect to time and place.* British journal of preventive & social medicine, 1973. **27**(4): p. 217-224.
30. Jaro, M.A. *UNIMATCH: a computer system for generalized record linkage under conditions of uncertainty.* in *Proceedings of the May 16-18, 1972, spring joint computer conference.* 1972. ACM.
31. Heasman, M. and J. Clarke, *Medical record linkage in Scotland.* Health bulletin, 1979. **37**(4): p. 97-103.
32. Black, D.W., G. Warrack, and G. Winokur, *The Iowa record-linkage study: I. Suicides and accidental deaths among psychiatric patients.* Archives of General Psychiatry, 1985. **42**(1): p. 71-75.
33. Snowdon, D.A., R.L. PHILLIPS, and W. Choi, *Diet, obesity, and risk of fatal prostate cancer.* American journal of epidemiology, 1984. **120**(2): p. 244-250.
34. Steinhoff, P.G., et al., *Women who obtain repeat abortions: a study based on record linkage.* Family planning perspectives, 1979: p. 30-38.
35. Mattsson, B. and A. Wallgren, *Completeness of the Swedish cancer register non-notified cancer cases recorded on death certificates in 1978.* Acta Oncologica, 1984. **23**(5): p. 305-313.
36. Koskenvuo, M., et al., *Incidence and prognosis of ischaemic heart disease with respect to marital status and social class. A national record linkage study.* Journal of epidemiology and community health, 1981. **35**(3): p. 192-196.
37. Martin, C., M. Hobbs, and B. Armstrong, *Identification of non-fatal myocardial infarction through hospital discharge data in Western Australia.* Journal of chronic diseases, 1987. **40**(12): p. 1111-1120.
38. Roos, L.L. and J.P. Nicol, *A research registry: uses, development, and accuracy.* Journal of clinical epidemiology, 1999. **52**(1): p. 39-47.
39. Holman, C.D.J., et al., *Population-based linkage of health records in Western Australia: development of a health services research linked database.* Australian and New Zealand Journal of Public Health, 1999. **23**(5): p. 453-459.
40. Kendrick, S. and J. Clarke, *The Scottish Record Linkage System.* Health bulletin, 1993. **51**(2): p. 72.
41. Ludvigsson, J.F., et al., *The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research.* European journal of epidemiology, 2009. **24**(11): p. 659-667.
42. Boyd, J.H., et al., *Technical challenges of providing record linkage services for research.* BMC medical informatics and decision making, 2014. **14**(1): p. 23.
43. Hobbs, M. and M. McCall, *Health statistics and record linkage in Australia.* Journal of chronic Diseases, 1970. **23**(5): p. 375-381.
44. Kelman, C., A. Bass, and D. Holman, *Research use of linked health data: A best practice protocol.* Australian and New Zealand Journal of Public Health, 2002. **26**: p. 5.
45. MacMahon, B. and T.F. Pugh, *Epidemiology: principles and methods.* Epidemiology: principles and methods., 1970.
46. Ahrens, W. and I. Pigeot, *Handbook of epidemiology.* 2005: Springer.
47. Rothman, K.J., S. Greenland, and T.L. Lash, *Modern epidemiology.* 2008: Lippincott Williams & Wilkins.

48. Sydney, T.U.o. *Introductory Analysis of Linked Data - Professional Development Course.* 2014; Available from: http://sydney.edu.au/medicine/public-health/future-student/study-program/professional-development/introduction-analysis-linked-data.php.

49. Michigan, U.o., *Introduction to Record Linkage.* 2014.

50. London, U.C. *Short Course: Record Linkage.* 2014; Available from: http://www.ucl.ac.uk/farr-short-courses/scfarr12.

51. Hobbs, M., et al., *The incidence of pneumoconiosis, mesothelioma and other respiratory cancer in men engaged in mining and milling crocidolite in Western Australia.* IARC scientific publications, 1979(30): p. 615-625.

52. Ferrante, A., *Developing an offender-based tracking system: The Western Australia INOIS project.* Australian & New Zealand Journal of Criminology, 1993. **26**(3): p. 232-250.

53. Stanley, F.J., et al., *A population database for maternal and child health research in Western Australia using record linkage.* Paediatric and perinatal epidemiology, 1994. **8**(4): p. 433-447.

54. Brook, E., et al., *Summary report: research outputs project, WA Data Linkage Unit (1995-2003).* population, 2004. **23**(5): p. 464-467.

55. Data Linkage Branch WA, *Submission for Productivity Commission Inquiry into Data Availability and Use.* 2016: Productivity Commission.

56. Linkage, C.f.H.R. *Master Linkage Key (MLK).* 2014 30 December 2014]; Available from: http://www.cherel.org.au/master-linkage-key.

57. Irvine, K.A. and L.K. Taylor, *The Centre for Health Record Linkage: fostering population health research in NSW.* New South Wales public health bulletin, 2011. **22**(2): p. 17-18.

58. Population Health Research Network. *Secure Unified Research Environment.* 2016 18/12/2016]; Available from: http://www.phrn.org.au/for-data-custodians/secure-unified-research-environment/.

59. Boyd, J.H., et al., *Data linkage infrastructure for cross-jurisdictional health-related research in Australia.* BMC health services research, 2012. **12**(1): p. 480.

60. Boyd, J.H., et al., *Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia.* BMC Health Services Research, 2015. **15**(1): p. 312.

61. Steorts, R.C., et al. *A comparison of blocking methods for record linkage.* in *Privacy in Statistical Databases.* 2014. Springer.

62. de Vries, T., et al. *Robust record linkage blocking using suffix arrays.* in *Proceedings of the 18th ACM conference on Information and knowledge management.* 2009. ACM.

63. Bachteler, T., J. Reiher, and R. Schnell, *Similarity filtering with multibit trees for record linkage.* 2013, Working Paper WP-GRLC-2013-02, German Record Linkage Center, Nuremberg.

64. Tromp, M., et al., *Ignoring dependency between linking variables and its impact on the outcome of probabilistic record linkage studies.* Journal of the American Medical Informatics Association, 2008. **15**(5): p. 654-660.

65. Zhu, V.J., et al., *An empiric modification to the probabilistic record linkage algorithm using frequency-based weight scaling.* Journal of the American Medical Informatics Association, 2009. **16**(5): p. 738-745.

66. Koudas, N., S. Sarawagi, and D. Srivastava. *Record linkage: similarity measures and algorithms.* in *Proceedings of the 2006 ACM SIGMOD international conference on Management of data.* 2006. ACM.

67. O'Keefe, C.M. and C.J. Connolly, *Privacy and the use of health data for research.* Medical Journal of Australia, 2010. **193**(9): p. 537-541.

68. Sethi, N. and G.T. Laurie, *Delivering proportionate governance in the era of eHealth: making linkage and privacy work together.* Medical law international, 2013: p. 0968533213508974.

69. Hernández, M.A. and S.J. Stolfo. *The merge/purge problem for large databases.* in *ACM SIGMOD Record.* 1995. ACM.

70. Jaro, M.A., *Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida.* Journal of the American Statistical Association, 1989. **89**: p. 414-420.

71. Brinkman, S., et al., *Associations between the early development instrument at age 5, and reading and numeracy skills at ages 8, 10 and 12: a prospective linked data study.* Child Indicators Research, 2013. **6**(4): p. 695-708.

72. Ferrante, A.M., *The Use of Data-Linkage Methods in Criminal Justice Research: A Commentary on Progress, Problems and Future Possibilities.* Current Issues in Criminal Justice, 2009. **20**(3): p. 1-15.

73. Stanley, P.F., *Developmental Pathways in WA Children Project.* 2006-2007, Chief Investigator and Director, Telethon Institute for Child Health Research: Perth WA.

74. Wong, C.X., et al., *Nationwide trends in the incidence of acute myocardial infarction in Australia, 1993–2010.* The American journal of cardiology, 2013. **112**(2): p. 169-173.

75. O'Flaherty, M., M.D. Huffman, and S. Capewell, *Declining trends in acute myocardial infarction attack and mortality rates, celebrating progress and ensuring future success.* Heart, 2015: p. heartjnl-2015-307868.

76. Smolina, K., et al., *Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study.* Bmj, 2012. **344**.

77. Jeschke, M.G., et al., *Long-term persistance of the pathophysiologic response to severe burn injury.* PloS one, 2011. **6**(7): p. e21245.

78. Duke, J.M., et al., *Mortality after burn injury in children: a 33-year population-based study.* Pediatrics, 2015. **135**(4): p. e903-e910.

79. Duke, J.M., et al., *Long-term mortality among older adults with burn injury: a population-based study in Australia.* Bulletin of the World Health Organization, 2015. **93**(6): p. 400-406.

80. Duke, J.M., et al., *Long-term Effects of Pediatric Burns on the Circulatory System.* Pediatrics, 2015. **136**(5): p. e1323-e1330.

81. Duke, J.M., et al., *Understanding the long-term impacts of burn on the cardiovascular system.* Burns, 2016. **42**(2): p. 366-374.

82. Randall, S.M., et al., *Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study.* BMJ open, 2015. **5**(9): p. e009395.

83. Churches, T., Christen, P., Lim, K., Zhu, J. X. , *Preparation of name and address data for record linkage using hidden Markov models.* BMC Medical Informatics and Decision Making, 2002. **2**(9).

84. Arasu, A., S. Chaudhuri, and R. Kaushik. *Transformation-based framework for record matching.* in *2008 IEEE 24th International Conference on Data Engineering.* 2008. IEEE.

85. Michelson, M. and C.A. Knoblock. *Mining heterogeneous transformations for record linkage.* in *Proceedings of the 6th International Workshop on Information Integration on the Web.* 2007.

86. Kumar, R. and R. Chadrasekaran, *Attribute correction-data cleaning using association rule and clustering methods.* Intl. Jrnl. of Data Mining & Knowledge Management Process, 2011. **1**(2): p. 22-32.

87. Ciszak, L. *Application of clustering and association methods in data cleaning.* in *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on.* 2008. IEEE.

88. Vick, R. and L. Huynh, *The effects of standardizing names for record linkage: Evidence from the United States and Norway.* Historical Methods, 2011. **44**(1): p. 15-24.

89. Grannis, S.J., J.M. Overhage, and C. McDonald, *Real world performance of approximate string comparators for use in patient matching.* Medinfo, 2004. **2004**: p. 43-7.

90. Li, X., et al. *Implementation of an extended Fellegi-Sunter probabilistic record linkage method using the Jaro-Winkler string comparator.* in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).* 2014. IEEE.

91. Cohen, W., P. Ravikumar, and S. Fienberg. *A comparison of string metrics for matching names and records.* in *Kdd workshop on data cleaning and object consolidation.* 2003.

92. Yancey, W.E., *Evaluating string comparator performance for record linkage.* Statistical Research Division Research Report, http://www. census. gov/srd/papers/pdf/rrs2005-05. pdf, 2005.

93. Snae, C., *A comparison and analysis of name matching algorithms.* International Journal of Applied Science. Engineering and Technology, 2007. **4**(1): p. 252-257.

94. Durham, E., et al., *Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage.* Information Fusion, 2012. **13**(4): p. 245-259.

95. Kılınç, D., *An accurate toponym-matching measure based on approximate string matching.* Journal of Information Science, 2015: p. 0165551515590097.

96. Lachance, M., *Useful functionalities for record linkage.*

97. Hermans, M. and F.C. Schadd. *A generalization of the winkler extension and its application for ontology mapping.* in *Proceedings Of The 24th Benelux Conference on Artificial Intelligence (BNAIC 2012).* 2012.

98. Bilenko, M., et al., *Adaptive name matching in information integration.* IEEE Intelligent Systems, 2003. **18**(5): p. 16-23.

99. Bilenko, M. and R.J. Mooney. *Adaptive duplicate detection using learnable string similarity measures.* in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* 2003. ACM.

100. Bilenko, M.Y., *Learnable similarity functions and their application to record linkage and clustering.* Vol. 67. 2006.

101. McCallum, A., K. Bellare, and F. Pereira, *A conditional random field for discriminatively-trained finite-state string edit distance.* arXiv preprint arXiv:1207.1406, 2012.

102. Schroeder, A.D. *Pad and Chaff: secure approximate string matching in private record linkage.* in *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services.* 2012. ACM.

103. Odell, K.M. and R.C. Russell, *Soundex phonetic comparison system*, U. Patents, Editor. 1918.

104. Taft, R.L., *Name Search Techniques*, in *New York State Identification and Intelligence System.* 1970: Albany, New York.

105. Winkler, W.E., *The State of Record Linkage and Current Research Problems.* U. S. Bureau of the Census.

106. Hettiarachchi, G.P. and D. Attygalle. *SPARCL: An improved approach for matching Sinhalese words and names in record clustering and linkage.* in *Global Humanitarian Technology Conference (GHTC), 2012 IEEE.* 2012. IEEE.

107. El-Shishtawy, T., *Linking Databases using Matched Arabic Names.* Computational Linguistics and Chinese Language Processing, 2014. **19**(1): p. 33-54.

108. del Pilar Angeles, M., A. Espino-Gamez, and J. Gil-Moncada. *Comparison of a Modified Spanish Phonetic, Soundex, and Phonex coding functions during data matching process.* in *Informatics, Electronics & Vision (ICIEV), 2015 International Conference on.* 2015. IEEE.

109. Miller, K.J. and M. Arehart. *Result Aggregation for Knowledge-Intensive Multicultural Name Matching.* in *Language and Technology Conference.* 2007. Springer.

110. Freeman, A.T., S.L. Condon, and C.M. Ackerman. *Cross linguistic name matching in English and Arabic: a one to many mapping extension of the Levenshtein edit distance algorithm.* in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.* 2006. Association for Computational Linguistics.

111. Lariscy, J.T., *Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox.* Journal of aging and health, 2011: p. 0898264311421369.

112. Herzog, T.N., F.J. Scheuren, and W.E. Winkler, *Estimating the Parameters of the Fellegi–Sunter Record Linkage Model*, in *Data Quality and Record Linkage Techniques.* 2007, Springer. p. 93-106.

113. McGlincy, M.H., *Using Test Databases to Evaluate Record Linkage Models and Train Linkage Practitioners.* Proceedings of the 29th American Statistical Association, Survey Research Method Section, Seattle, WA, 2006: p. 3404-3410.

114. Verykios, V.S., et al. *On the Accuracy and Completeness of the Record Matching Process.* in *IQ.* 2000.

115. McDonald, C.J., *Analysis of a probabilistic record linkage technique without human review.* 2003.

116. Bauman, G.J., *Computation of weights for probabilistic record linkage using the EM algorithm.* 2006.

117. Yancey, W.E., *Improving EM algorithm estimates for record linkage parameters.* Proceedings of the Section on Survey Research Methods, American Statisitcal Association, 2002.

118. Zhu, Y., et al., *Comparative validity of methods to select appropriate cutoff weight for probabilistic linkage without unique personal identifiers.* Pharmacoepidemiology and drug safety, 2015.

119. Lahiri, P. and M.D. Larsen, *Regression analysis with linked data.* Journal of the American statistical association, 2005. **100**(469): p. 222-230.

120. Larsen, M.D. and D.B. Rubin, *Iterative automated record linkage using mixture models.* Journal of the American Statistical Association, 2001. **96**(453): p. 32-41.

121. Newcombe, H.B., *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business.* 1988, New York: Oxford University Press.

122. Daggy, J., et al., *Evaluating latent class models with conditional dependence in record linkage.* Statistics in medicine, 2014. **33**(24): p. 4250-4265.

123. Zhu, R., et al., *Stepwise Variable Selection for Loglinear Mixture in Record Linkage.* European Journal of Pure and Applied Mathematics, 2010. **3**(2): p. 141-162.

124. Porter, E.H. and W.E. Winkler. *Approximate string comparison and its effect on an advanced record linkage system.* in *Advanced record linkage system. US Bureau of the Census, Research Report.* 1997. Citeseer.

125. Li, X., et al. *An empiric weight computation for record linkage using linearly combined fields' similarity scores.* in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* 2014. IEEE.

126. Winkler, W.E., *Frequency-Based Matching in Fellegi-Sunter Model of Record Linkage.* 2000/06.

127. Yancey, W.E., *Frequency-Dependent Probability Measures for Record Linkage.* 2000/07.

128. Bhattacharya, I. and L. Getoor. *Iterative record linkage for cleaning and integration.* in *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery.* 2004. ACM.

129. Domingos, P. *Multi-relational record linkage.* in *In Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining.* 2004. Citeseer.

130. Guo, S., et al., *Record linkage with uniqueness constraints and erroneous values.* Proceedings of the VLDB Endowment, 2010. **3**(1-2): p. 417-428.

131. On, B.-W., et al. *Group linkage.* in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on.* 2007. IEEE.

132. Kendrick, S., et al., *Best-link matching of Scottish health data sets.* Methods of information in medicine, 1998. **37**(1): p. 64.

133. MacLeod, M.C., et al., *Enhancing the power of record linkage involving low quality personal identifiers: use of the best link principle and cause of death prior likelihoods.* Computers and biomedical research, 1998. **31**(4): p. 257-270.

134. Chen, C., et al. *Methodology for Large-Scale Entity Resolution without Pairwise Matching*. in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. 2015. IEEE.

135. Larsen, M.D. *Modeling issues and the use of experience in record linkage*. in *Record Linkage Techniques-1997: Proceedings of an International Workshop and Exposition*. 1997. Citeseer.

136. Fortini, M., et al., *On Bayesian record linkage*. Research in Official Statistics, 2001. **4**(1): p. 185-198.

137. Zhou, Y., et al. *Extending naive bayes classifier with hierarchy feature level information for record linkage*. in *Workshop on Advanced Methodologies for Bayesian Networks*. 2015. Springer.

138. Fortini, M., et al. *Modelling issues in record linkage: a Bayesian perspective*. in *Proceedings of the American Statistical Association, Survey Research Methods Section*. 2002.

139. Elkany, A.E.M.C.P. *An efficient domain-independent algorithm for detecting approximately duplicate database records*. in *Proc of the SIGMOD*.

140. Cheong, Y. and J. Tay. *Approximate string matching for multiple-attribute, large-scale customer address databases*. in *International Conference on Asian Digital Libraries*. 2003. Springer.

141. Bilenko, M. and R.J. Mooney, *Learning to combine trained distance metrics for duplicate detection in databases*. Submitted to CIKM-2002, 2002: p. 1-19.

142. Borthwick, A., M. Buechi, and A. Goldberg. *Key concepts in the choicemaker 2 record matching system*. in *Procs. First Workshop on Data Cleaning, Record Linkage, and Object Consolidation, in conjunction with KDD*. 2003.

143. Hettiarachchi, G.P., et al. *Next generation data classification and linkage: Role of probabilistic models and artificial intelligence*. in *Global Humanitarian Technology Conference (GHTC), 2014 IEEE*. 2014. IEEE.

144. Ventura, S.L. and R. Nugent. *Hierarchical Linkage Clustering with Distributions of Distances for Large-Scale Record Linkage*. in *International Conference on Privacy in Statistical Databases*. 2014. Springer.

145. De Leone, R. and V. Minnetti, *Electre Tri-Machine Learning Approach to the Record Linkage Problem*. arXiv preprint arXiv:1505.06614, 2015.

146. Pernelle, N. and F. Saïs. *Classification rule learning for data linking*. in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*. 2012. ACM.

147. Paskalev, P. and V. Nikolov. *Evaluation of records similarity in a duplication search engine using neural network*. in *2008 4th International IEEE Conference Intelligent Systems*. 2008.

148. Christen, P. *Automatic record linkage using seeded nearest neighbour and support vector machine classification*. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2008. ACM.

149. Christen, P. *A two-step classification approach to unsupervised record linkage*. in *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*. 2007. Australian Computer Society, Inc.

150. Gu, L. and R. Baxter. *Decision models for record linkage*. in *Data Mining*. 2006. Springer.

151. Sariyar, M. and A. Borg, *Bagging, bumping, multiview, and active learning for record linkage with empirical results on patient identity data*. Computer methods and programs in biomedicine, 2012. **108**(3): p. 1160-1169.

152. Ravikumar, P. and W.W. Cohen. *A hierarchical graphical model for record linkage*. in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. 2004. AUAI Press.

153. Michalowski, M., S. Thakkar, and C.A. Knoblock, *Exploiting secondary sources for unsupervised record linkage*. 2004, DTIC Document.

154. Sariyar, M., A. Borg, and K. Pommerening, *Evaluation of record linkage methods for iterative insertions*. Methods of information in medicine, 2009. **48**(5): p. 429-437.

155. Ektefa, M., et al., *A comparative study in classification techniques for unsupervised record linkage model.* Journal of Computer Science, 2011. **7**(3): p. 341.

156. Christen, P. *Automatic training example selection for scalable unsupervised record linkage.* in *Pacific-Asia Conference on Knowledge Discovery and Data Mining.* 2008. Springer.

157. Wang, Q., D. Vatsalan, and P. Christen. *Efficient interactive training selection for large-scale entity resolution.* in *Pacific-Asia Conference on Knowledge Discovery and Data Mining.* 2015. Springer.

158. Arasu, A., M. Götz, and R. Kaushik. *On active learning of record matching packages.* in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data.* 2010. ACM.

159. Bilenko, M., S. Basil, and M. Sahami. *Adaptive product normalization: Using online learning for record linkage in comparison shopping.* in *Fifth IEEE International Conference on Data Mining (ICDM'05).* 2005. IEEE.

160. Dong, X., A. Halevy, and J. Madhavan. *Reference reconciliation in complex information spaces.* in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data.* 2005. ACM.

161. On, B.-W., et al. *Comparative study of name disambiguation problem using a scalable blocking-based framework.* in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries.* 2005. ACM.

162. Dredze, M., et al. *Entity disambiguation for knowledge base population.* in *Proceedings of the 23rd International Conference on Computational Linguistics.* 2010. Association for Computational Linguistics.

163. Yakout, M., et al., *Behavior based record linkage.* Proceedings of the VLDB Endowment, 2010. **3**(1-2): p. 439-448.

164. Malin, B., E. Airoldi, and K.M. Carley, *A network analysis model for disambiguation of names in lists.* Computational & Mathematical Organization Theory, 2005. **11**(2): p. 119-139.

165. Singla, P. and P. Domingos. *Entity resolution with markov logic.* in *Sixth International Conference on Data Mining (ICDM'06).* 2006. IEEE.

166. Shen, W., X. Li, and A. Doan. *Constraint-based entity matching.* in *AAAI.* 2005.

167. Rehman, M. and V. Esichaikul. *Duplicate record detection for database cleansing.* in *Machine Vision, 2009. ICMV'09. Second International Conference on.* 2009. IEEE.

168. Köpcke, H., A. Thor, and E. Rahm, *Evaluation of entity resolution approaches on real-world match problems.* Proceedings of the VLDB Endowment, 2010. **3**(1-2): p. 484-493.

169. Gibbs, T.H., *A declarative approach to entity resolution*, in *Data Engineering.* 2009, Springer. p. 17-38.

170. Syed, H., et al. *Evaluation of Entity Resolution Results through Benchmarking and Truth set Development.* in *Proceedings of the International Conference on Information and Knowledge Engineering (IKE).* 2013. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

171. Clark, D.E. and D.R. Hahn. *Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry.* in *Proceedings of the Annual Symposium on Computer Application in Medical Care.* 1995. American Medical Informatics Association.

172. Tromp, M., et al., *Medical record linkage of anonymous registries without validated sample linkage of the Dutch perinatal registries.* Studies in health technology and informatics, 2005. **116**: p. 125-130.

173. Zhu, Y., et al., *When to conduct probabilistic linkage vs. deterministic linkage? A simulation study.* Journal of biomedical informatics, 2015. **56**: p. 80-86.

174. Tromp, M., et al., *Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage.* Journal of clinical epidemiology, 2011. **64**(5): p. 565-572.

175. Rotermann, M., et al., *Two approaches to linking census and hospital data.* Health reports, 2014. **25**(10): p. 3.

176. Ferrante, A. and J. Boyd, *A transparent and transportable methodology for evaluating Data Linkage software.* Journal of Biomedical Informatics, 2012. **45**(1): p. 165-172.
177. Campbell, K.M., *Impact of record-linkage methodology on performance indicators and multivariate relationships.* Journal of substance abuse treatment, 2009. **36**(1): p. 110-117.
178. Baldi, I., et al., *The impact of record-linkage bias in the Cox model.* Journal of evaluation in clinical practice, 2010. **16**(1): p. 92-96.
179. Brenner, H., I. Schmidtmann, and C. Stegmaier, *Effects of record linkage errors on registry-based follow-up studies.* Statistics in Medicine, 1997. **16**(23): p. 2633-2643.
180. Oberaigner, W., *Errors in survival rates caused by routinely used deterministic record linkage methods.* Methods Inf Med, 2007. **46**(4): p. 420-424.
181. Bohensky, M.A., et al., *Data linkage: a powerful research tool with potential problems.* BMC health services research, 2010. **10**(1): p. 346.
182. Chambers, R., *Regression analysis of probability-linked data.* Official Statistics Research Series, 2009. **4**(2).
183. Kim, G. and R. Chambers, *Regression analysis under incomplete linkage.* Computational Statistics & Data Analysis, 2012. **56**(9): p. 2756-2770.
184. Kim, G. and R. Chambers, *Regression Analysis under Probabilistic Multi-Linkage.* Statistica Neerlandica, 2012. **66**(1): p. 64-79.
185. Hof, M. and A. Zwinderman, *Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables.* Statistics in medicine, 2012. **31**(30): p. 4231-4242.
186. Goldstein, H., K. Harron, and A. Wade, *The analysis of record-linked data using multiple imputation with data value priors.* Statistics in medicine, 2012. **31**(28): p. 3481-3493.
187. Harron, K., et al., *Evaluating bias due to data linkage error in electronic healthcare records.* BMC medical research methodology, 2014. **14**(1): p. 36.
188. Warke, Y., A. Mohanpurkar, and S. Phule, *Contraption of Suffix Array Blocking for Efficacious Record Linkage and De-duplication.*
189. Yan, S., et al. *Adaptive sorted neighborhood methods for efficient record linkage.* in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.* 2007. ACM.
190. Draisbach, U. and F. Naumann. *A comparison and generalization of blocking and windowing algorithms for duplicate detection.* in *Proceedings of the International Workshop on Quality in Databases (QDB).* 2009.
191. Rendle, S. and L. Schmidt-Thieme. *Scaling record linkage to non-uniform distributed class sizes.* in *Pacific-Asia Conference on Knowledge Discovery and Data Mining.* 2008. Springer.
192. Shin, J., *Comparative Study On Blocking Methods In Record Linkage.* 2009.
193. Sarma, A.D., et al., *CBLOCK: An Automatic Blocking Mechanism for Large-Scale De-duplication Tasks.* arXiv preprint arXiv:1111.3689, 2011.
194. Bilenko, M., B. Kamath, and R.J. Mooney. *Adaptive blocking: Learning to scale up record linkage.* in *Data Mining, 2006. ICDM'06. Sixth International Conference on.* 2006. IEEE.
195. Prasad, K.H., et al. *Automated selection of blocking columns for record linkage.* in *Service Operations and Logistics, and Informatics (SOLI), 2012 IEEE International Conference on.* 2012. IEEE.
196. Giang, P.H., *A machine learning approach to create blocking criteria for record linkage.* Health care management science, 2015. **18**(1): p. 93-105.
197. Ramadan, B. and P. Christen. *Unsupervised blocking key selection for real-time entity resolution.* in *Pacific-Asia Conference on Knowledge Discovery and Data Mining.* 2015. Springer.
198. Michelson, M. and C.A. Knoblock. *Learning blocking schemes for record linkage.* in *Proceedings of the National Conference on Artificial Intelligence.* 2006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

199. Quantin, C., et al., *Which are the best identifiers for record linkage?* Medical informatics and the Internet in medicine, 2004. **29**(3-4): p. 221-227.

200. Randall, S.M., et al., *Privacy-preserving record linkage on large real world datasets.* Journal of biomedical informatics, 2014. **50**: p. 205-212.

201. Schnell, R., *Privacy-preserving record linkage and privacy-preserving blocking for large files with cryptographic keys using multibit trees.* JSM, 2013.

202. Karakasidis, A. and V.S. Verykios. *A highly efficient and secure multidimensional blocking approach for private record linkage.* in *2012 IEEE 24th International Conference on Tools with Artificial Intelligence.* 2012. IEEE.

203. Ranbaduge, T., D. Vatsalan, and P. Christen. *Clustering-based scalable indexing for multi-party privacy-preserving record linkage.* in *Pacific-Asia Conference on Knowledge Discovery and Data Mining.* 2015. Springer.

204. Karapiperis, D. and V.S. Verykios, *An LSH-based blocking approach with a homomorphic matching technique for privacy-preserving record linkage.* IEEE Transactions on Knowledge and Data Engineering, 2015. **27**(4): p. 909-921.

205. Karakasidis, A., G. Koloniari, and V.S. Verykios. *Scalable blocking for privacy preserving record linkage.* in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2015. ACM.

206. Ford, D.V., et al., *The SAIL Databank: building a national architecture for e-health research and evaluation.* 4 September 2009.

207. Jutte, D.P., L.L. Roos, and M.D. Brownell, *Administrative record linkage as a tool for public health research.* Annual review of public health, 2011. **32**: p. 91-108.

208. Quantin, C., et al., *How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure.* International journal of medical informatics, 1998. **49**(1): p. 117-122.

209. Al-Lawati, A., D. Lee, and P. McDaniel. *Blocking-aware private record linkage.* in *Proceedings of the 2nd international workshop on Information quality in information systems.* 2005. ACM.

210. Ravikumar, P., W.W. Cohen, and S.E. Fienberg, *A secure protocol for computing string distance metrics.* PSDM held at ICDM, 2004: p. 40-46.

211. Inan, A., et al. *A hybrid approach to private record linkage.* in *2008 IEEE 24th International Conference on Data Engineering.* 2008. IEEE.

212. Kuzu, M., et al. *Efficient privacy-aware record integration.* in *Proceedings of the 16th International Conference on Extending Database Technology.* 2013. ACM.

213. Inan, A., et al. *Private record matching using differential privacy.* in *Proceedings of the 13th International Conference on Extending Database Technology.* 2010. ACM.

214. Scannapieco, M., et al. *Privacy preserving schema and data matching.* in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data.* 2007. ACM.

215. Yakout, M., M.J. Atallah, and A. Elmagarmid, *Efficient and practical approach for private record linkage.* Journal of Data and Information Quality (JDIQ), 2012. **3**(3): p. 5.

216. Pang, C. and D. Hansen, *Improved record linkage for encrypted identifying data.* HIC 2006 and HINZ 2006: Proceedings, 2006: p. 164.

217. Pang, C., et al., *Privacy-preserving fuzzy matching using a public reference table*, in *Intelligent Patient Management.* 2009, Springer. p. 71-89.

218. Thoben, W., H.-J. Appelrath, and S. Sauer, *Record linkage of anonymous data by control numbers*, in *From Data to Knowledge.* 1996, Springer. p. 412-419.

219. Karmel, R., et al., *Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study.* 2010.

220. Boyle, D.I.R. and N. Rafael, *BioGrid Australia and GRHANITE: Privacy-Protecting Subject Matching* Studies in Health Technology and Informatics 2011. **168**: p. 24-34.

221. Van Eycken, E., et al., *Evaluation of the encryption procedure and record linkage in the Belgian National Cancer Registry.* Archives of public health, 2000. **58**(6): p. 281-294.

222. Boyd, J.H., S.M. Randall, and A.M. Ferrante, *Application of Privacy-Preserving Techniques in Operational Record Linkage Centres*, in *Medical Data Privacy Handbook.* 2015, Springer. p. 267-287.

223. Karakasidis, A. and V.S. Verykios, *Secure Blocking+ Secure Matching= Secure Record Linkage.* JCSE, 2011. **5**(3): p. 223-235.

224. Vatsalan, D. and P. Christen. *An iterative two-party protocol for scalable privacy-preserving record linkage.* in *Proceedings of the Tenth Australasian Data Mining Conference-Volume 134.* 2012. Australian Computer Society, Inc.

225. Vatsalan, D., P. Christen, and V.S. Verykios. *An efficient two-party protocol for approximate matching in private record linkage.* in *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121.* 2011. Australian Computer Society, Inc.

226. Durham, E., et al. *Private medical record linkage with approximate matching.* in *AMIA Annual Symposium Proceedings.* 2010. American Medical Informatics Association.

227. Durham, E.A., *A framework for accurate, efficient private record linkage.* 2012, University of Texas at Dallas.

228. Durham, E.A., et al., *Composite bloom filters for secure record linkage.* IEEE transactions on knowledge and data engineering, 2014. **26**(12): p. 2956-2968.

229. Schnell, R., T. Bachteler, and J. Reiher, *A Novel Error-Tolerant Anonymous Linking Code.* 2011, Working Paper Series No. WP-GRLC-2011-02. Nürnberg, Germany: German Record Linkage Center.

230. Niedermeyer, F., et al., *Cryptanalysis of basic Bloom filters used for privacy preserving record linkage.* Journal of Privacy and Confidentiality, 2014. **6**(2): p. 3.

231. Kroll, M. and S. Steinmetzer, *Automated Cryptanalysis of Bloom Filter Encryptions of Health Records.* arXiv preprint arXiv:1410.6739, 2014.

232. Makel, M.C., J.A. Plucker, and B. Hegarty, *Replications in psychology research how often do they really occur?* Perspectives on Psychological Science, 2012. **7**(6): p. 537-542.

233. Matosin, N., et al., *Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture.* Disease Models and Mechanisms, 2014. **7**(2): p. 171-173.

234. Etzkowitz, H., et al., *The future of the university and the university of the future: evolution of ivory tower to entrepreneurial paradigm.* Research policy, 2000. **29**(2): p. 313-330.

235. Borins, S., *Encouraging innovation in the public sector.* Journal of intellectual capital, 2001. **2**(3): p. 310-319.

236. Boyd, J.H., et al., *Understanding the origins of record linkage errors and how they affect research outcomes.* Australian and New Zealand Journal of Public Health, 2016.

237. Christen, P. and K. Goiser, *Quality and Complexity Measures for Data Linkage and Deduplication*, in *Quality Measures in Data Mining.* 2007. p. 127-151.

238. Boyd, J., et al., *A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects.* Methods of information in medicine, 2016. **55**(3): p. 276-283.

239. Christen, P., T. Churches, and M. Hegland, *Febrl–a parallel open source data linkage system*, in *Advances in Knowledge Discovery and Data Mining.* 2004, Springer. p. 638-647.

240. *Linkage Wiz Data Matching Software* [cited 2013 7th May]; Available from: http://www.linkagewiz.net/.

241. Randall, S.M., et al., *The effect of data cleaning on record linkage quality.* BMC Medical Informatics and Decision Making, 2013. **13**(1): p. 64.

242. Christen, P., *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.* 2012: Springer.

243. Randall, S.M., et al., *Grouping methods for ongoing record linkage.* Population Informatics for Big Data, Sydney, Australia, 2015.

244. Randall, S.M., et al., *Use of graph theory measures to identify errors in record linkage.* Computer methods and programs in biomedicine, 2014. **115**(2): p. 55-63.

245. Al-Shahi, R. and C. Warlow, *Using patient-identifiable data for observational research and audit.* BMJ, 2000. **321**(7268): p. 1031-1032.

246. O'Keefe, C.M. and C. Connolly, *Regulation and perception concerning the use of health data for research in australia.* electronic Journal of Health Informatics, 2011. **6**(2): p. 16.
247. Holman, C.D.A.J., *Anonymity and Research: Health Data and Biospecimen Law in Australia.* 2012: Uniprint, UWA.
248. *Privacy Act* 1988, http://www.comlaw.gov.au/Details/C2013C00482: Australia.
249. National Health and Medical Research Council, *Guidelines under Section 95 of the Privacy Act 1988.* 2014: http://www.nhmrc.gov.au/guidelines/publications/pr1.
250. Lovett, R., et al., *A review of Australian health privacy regulation regarding the use and disclosure of identified data to conduct data linkage.* 2006.
251. Health, N. and M.R. Council, *National statement on ethical conduct in human research.* 2007: National Health and Medical Research Council.
252. NAB, *NAB Charitable Giving Index.* 2016.
253. Office of the Australian Information Commissioner, *Community attitudes to privacy survey, Research report.* 2013.
254. Office of the Privacy Commissioner; Australia, *Community Attitudes to Privacy.* 2007.
255. Office of the Federal Privacy Commissioner, *Community Attitudes Towards Privacy 2004.* 2004.
256. Australian Medical Association, *AMA poll shows patients are concerned about the privacy and security of their medical records. Canberra: AMA, 2005.*
257. Cabinet Secretary, S.H.J.L., *Australian Law Reform Commission Report 108 For Your Information: Australian Privacy Law and Practice.* October 2009.
258. The Consumers' Health Forum of Australia, *Towards the best use of personal health information - a consumer perspective.* 1998: https://chf.org.au/.
259. Jones, K.H., C.L. McNerney, and D.V. Ford, *Involving consumers in the work of a data linkage research unit.* International Journal of Consumer Studies, 2014. **38**(1): p. 45-51.
260. Office of the Australian Information Commissioner, *Open public sector information: from principles to practice.* 2013.
261. Pullman, D., et al., *Sorry, you can't have that information: data holder confusion regarding privacy requirements for personal health information and the potential chilling effect on health research.* Healthcare Policy, 2009. **4**(4): p. 61.
262. Randall, S.M., et al., *Limited privacy protection and poor sensitivity Is it time to move on from the statistical linkage key-581?* Health Information Management Journal, 2016: p. 1833358316647587.
263. Randall, S.M., et al., *Privacy preserving record linkage using homomorphic encryption.* Population Informatics for Big Data, Sydney, Australia, 2015.
264. Centre for Health Record Linkage. *Achievements.* 2016 3/12/2016]; Available from: http://www.cherel.org.au/achievements.
265. Medical Research Future Fund, *Australian Medical Research and Innovation Strategy 2016-2021,.* 2016.
266. Office for National Statistics, *The Census and Future Provision of Population Statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority.* 2014.
267. Statistics New Zealand, *An overview of progress on the potential use of administrative data for census information in New Zealand* 2014.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Appendix A: Contributions to manuscripts – acknowledgements by co-authors

Boyd, J. H., Randall, S. M., Ferrante, A. M., Bauer, J. K., Brown, A. P., & Semmens, J. B. (2014). Technical challenges of providing record linkage services for research. *BMC medical informatics and decision making, 14*(1), 1.

Contribution:

SR developed the concepts within this paper in collaboration with the other co-authors, and assisted in writing the first draft of the manuscript.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Anna Ferrante: A Ferrante

Jacqui Bauer: Jacqueline Bauer

Adrian Brown:

James Semmens:

Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., Randall, S. M., & Semmens, J. B. (2012). Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC health services research*, *12*(1), 1

Contribution:

SR supported the development of this paper and made contributions to the final version of the manuscript.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Anna Ferrante: A Ferrante

Christine O'Keefe: ______________________

John Bass: ______________________

James Semmens: ______________________

Boyd, J. H., Randall, S. M., Ferrante, A. M., Bauer, J. K., McInneny, K., Brown, A. P., Spilsbury, K., Gillies, M, & Semmens, J. B. (2015). Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia. *BMC health services research*, *15*(1), 1.

Contribution:

SR supported the development of this paper and made contributions to the final version of the manuscript.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Anna Ferrante: A. Ferrante

Jacqui Bauer: Jacqueline Bauer

Kevin McInneny:

Adrian Brown:

Katrina Spilsbury: K.

Margo Gillies: M. Gillies

James Semmens:

Randall, S. M., Zilkens, R., Duke, J. M., & Boyd, J. H. (2016). Western Australia population trends in the incidence of acute myocardial infarction between 1993 and 2012. *International Journal of Cardiology, 222*, 678-682.

Contribution:

SR developed the methodology and research design, reviewed the literature, performed all analyses, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Renate Zilkens: RZ

Janine Duke: Janine Duke

James Boyd: James H Boyd

Duke, J. M., Rea, S., Boyd, J. H., Randall, S. M., & Wood, F. M. (2015). Mortality after burn injury in children: a 33-year population-based study. *Pediatrics, 135*(4), e903-e910.

Contribution:

SR supported the development of this paper, conducting the analysis with other co-authors, and making contributions to the final version of the manuscript.

I acknowledge the above statement of contribution is accurate:

Janine Duke: Janine Duke

Suzanne Rea: SRea

James Boyd: James H Boyd

Fiona Wood:

Duke, J. M., Boyd, J. H., Rea, S., Randall, S. M., & Wood, F. M. (2015). Long-term mortality among older adults with burn injury: a population-based study in Australia. *Bulletin of the World Health Organization, 93*(6), 400-406

Contribution:

SR supported the development of this paper, conducting the analysis with other co-authors, and making contributions to the final version of the manuscript.

I acknowledge the above statement of contribution is accurate:

Janine Duke: Janine Duke

James Boyd: James H Boyd

Suzanne Rea:

Fiona Wood:

Randall, S. M., Fear, M. W., Wood, F. M., Rea, S., Boyd, J. H., & Duke, J. M. (2015). Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study. *BMJ open*, *5*(9), e009395.

Contribution:

SR supported the development of this paper, conducting the analysis, interpreting results, writing the result section of the manuscript, and making contributions to the final version of the manuscript.

I acknowledge the above statement of contribution is accurate:

Mark Fear: ______________________

Fiona Wood: ______________________

Suzanne Rea: ______________________

James Boyd: James H Boyd

Janine Duke: Janine Duke

Boyd, J. H., Ferrante, A. M., Irvine, K., Smith, M., Moore, E., Brown, A. P., Randall, S. M. (2016). Understanding the origins of record linkage errors and how they affect research outcomes. *Australian and New Zealand Journal of Public Health*, In Press

Contribution:

SR supported the development of this paper, assisting with writing the first draft of the manuscript and contributing to the final version.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Anna Ferrante: A. Ferrante

Katie Irvine: Katie Irvine

Michael Smith: M Smith

Elizabeth Moore: Moore

Adrian Brown:

Boyd, J. H., Guiver, T., Randall, S. M., Ferrante, A. M., Semmens, J. B., Anderson, P., & Dickinson, T. (2016). A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. *Methods of information in medicine*, *55*(3), 276-283.

Contribution:

SR supported the development of this paper, collecting and analysing data, and contributing to the final version of the manuscript.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Tenniel Guiver: ______________________

Anna Ferrante: a. Ferrante

James Semmens: ______________________

Phil Anderson: ______________________

Teresa Dickinson: ______________________

Boyd, J. H., Guiver, T., Randall, S. M., Ferrante, A. M., Semmens, J. B., Anderson, P., & Dickinson, T. (2016). A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. *Methods of information in medicine, 55*(3), 276-283.

Contribution:

SR supported the development of this paper, collecting and analysing data, and contributing to the final version of the manuscript.

I acknowledge the above statement of contribution is accurate:

James Boyd: ______________________________

Tenniel Guiver: ______________________________

Anna Ferrante: ______________________________

James Semmens: ______________________________

Phil Anderson: ______________________________

Teresa Dickinson: ______________________________

Randall, S. M., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making, 13*(1), 64.

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, performed all evaluations, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Anna Ferrante: A. Ferrante

James Boyd: James H Boyd

James Semmens:

Randall, S. M., Boyd, J. H., Ferrante, A. M., Brown, A. P., Semmens, J. B. (2015). Grouping methods for ongoing record linkage. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, performed all analyses, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Anna Ferrante: A. Ferrante

Adrian Brown:

James Semmens:

Randall, S. M., Boyd, J. H., Ferrante, A. M., Bauer, J. K., & Semmens, J. B. (2014). Use of graph theory measures to identify errors in record linkage. *Computer methods and programs in biomedicine, 115*(2), 55-63.

Contribution:

SR developed the research design and evaluation methodology for the paper, developed and selected the graph theory measures used, reviewed the literature, performed all analyses, produced and interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Anna Ferrante: A. Ferrante

Jacqui Bauer: Jacqui Bauer

James Semmens:

Boyd, J. H., Randall, S. M., & Ferrante, A. M. (2015). Application of Privacy-Preserving Techniques in Operational Record Linkage Centres. In *Medical Data Privacy Handbook* (pp. 267-287). Springer International Publishing.

Contribution:

SR supported the development of this paper, assisting in writing the first draft of the manuscript and contributing to the final version.

I acknowledge the above statement of contribution is accurate:

James Boyd: James H Boyd

Anna Ferrante: A Ferrante

Randall, S. M., Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics, 50,* 205-212

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, developed the required software for linkage, performed all analyses, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Anna Ferrante: A. Ferrante

James Boyd: James H Boyd

Jacqui Bauer: Jacqueline Bauer

James Semmens:

Randall, S. M., Ferrante, A. M., Boyd, J. H., Brown, A. P., Semmens, J. B. (2016). Limited privacy protection and poor sensitivity: is it time to move on from the Statistical Linkage Key-581? *Health information management journal 45*(2), 71-79.

Contribution:

SR developed the research design and evaluation methodology for the paper, reviewed the literature, performed all evaluations, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Anna Ferrante: A Ferrante

James Boyd: James H Boyd

Adrian Brown:

James Semmens:

Randall, S. M., Brown, A. P., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2015). Privacy preserving record linkage using homomorphic encryption. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* Sydney, Australia.

Contribution:

SR developed the initial concept, formulated the research design and evaluation methodology for the paper, reviewed the literature, developed the required software, performed all evaluations, interpreted results, wrote the first draft of the manuscript, and edited the manuscript into its final form with the comments and suggestions of the other authors.

I acknowledge the above statement of contribution is accurate:

Adrian Brown: ____________________

Anna Ferrante: A Ferrante

James Boyd: James H Boyd

James Semmens: ____________________

Appendix B: Permissions to reproduce manuscripts

1. Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., Brown, A. P., & Semmens, J. B. (2014). Technical challenges of providing record linkage services for research. *BMC medical informatics and decision making, 14*(1), 1.

As mentioned in the footer of page 1 of this article, this articles is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Boyd, J. H., Ferrante, A. M., O'Keefe, C. M., Bass, A. J., **Randall, S. M.**, & Semmens, J. B. (2012). Data linkage infrastructure for cross-jurisdictional health-related research in Australia. *BMC health services research*, 12(1), 1.

As mentioned in the footer of page 1 of this article, this articles is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

3. Boyd, J. H., **Randall, S. M.**, Ferrante, A. M., Bauer, J. K., McInneny, K., Brown, A. P., Spilsbury, K., Gillies, M, & Semmens, J. B. (2015). Accuracy and completeness of patient pathways–the benefits of national data linkage in Australia. *BMC health services research*, *15*(1), 1.

As mentioned in the footer of page 1 of this article, this articles is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

4. **Randall, S. M.**, Zilkens, R., Duke, J. M., & Boyd, J. H. (2016). Western Australia population trends in the incidence of acute myocardial infarction between 1993 and 2012. *International Journal of Cardiology, 222*, 678-682.

A license to include this paper in this thesis has been added below.

5. Duke, J. M., Rea, S., Boyd, J. H., **Randall, S. M.**, & Wood, F. M. (2015). Mortality after burn injury in children: a 33-year population-based study. *Pediatrics, 135(*4), e903-e910.

A license to include this paper in this thesis has been added below.

6. Duke, J. M., Boyd, J. H., Rea, S., **Randall, S. M.**, & Wood, F. M. (2015). Long-term mortality among older adults with burn injury: a population-based study in Australia. *Bulletin of the World Health Organization, 93*(6), 400-406.

The publisher provides permission for a publication to be used in a thesis or dissertation without seeking permission, as long as the WHO source is appropriately acknowledged. Evidence of this can be found at

http://www.who.int/about/licensing/extracts/en/ (accessed 2 December 2016)

7. **Randall, S. M.**, Fear, M. W., Wood, F. M., Rea, S., Boyd, J. H., & Duke, J. M. (2015). Long-term musculoskeletal morbidity after adult burn injury: a population-based cohort study. *BMJ open*, *5*(9), e009395.

As mentioned on page 9 of this manuscript (Open Access), this article is distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits unrestricted distribution for non-commercial use, provided the original work is properly cited.

8. Boyd, J. H., Ferrante, A. M., Irvine, K., Smith, M., Moore, E., Brown, A. P., **Randall, S. M.** (2016). Understanding the origins of record linkage errors and how they affect research outcomes. *Australian and New Zealand Journal of Public Health, In Press*

A license to include this paper in this thesis has been added below.

9. Boyd, J. H., Guiver, T., **Randall, S. M.**, Ferrante, A. M., Semmens, J. B., Anderson, P., & Dickinson, T. (2016). A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. *Methods of information in medicine*, *55*(3), 276-283.

The publisher provides permission for an author to include the article in a thesis or dissertation, provided that this is not published commercially. Evidence of this can be found at

http://methods.schattauer.de/fileadmin/assets/zeitschriften/methods/Instructions_to_Authors_20160712.pdf

(Schattauer Copyright Permission Policy, pg 2; accessed 2 December 2016)

10. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. *BMC medical informatics and decision making, 13*(1), 64.

As mentioned in the footer of page 1 of the article, this articles is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

11. **Randall, S. M.**, Boyd, J. H., Ferrante, A. M., Brown, A. P., Semmens, J. B. (2015). Grouping methods for ongoing record linkage. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

As mentioned in the footer of page 1 of the article, the copyright of this article remains with the authors.

12. **Randall, S. M.**, Boyd, J. H., Ferrante, A. M., Bauer, J. K., & Semmens, J. B. (2014). Use of graph theory measures to identify errors in record linkage. *Computer methods and programs in biomedicine, 115*(2), 55-63.

A license to include this paper in this thesis has been added below.

13. Boyd, J. H., **Randall, S. M.**, & Ferrante, A. M. (2015). Application of Privacy-Preserving Techniques in Operational Record Linkage Centres. In *Medical Data Privacy Handbook* (pp. 267-287). Springer International Publishing.

A license to include this paper in this thesis has been added below.

14. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., Bauer, J. K., & Semmens, J. B. (2014). Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics, 50*, 205-212

A license to include this paper in this thesis has been added below.

15. **Randall, S. M.**, Ferrante, A. M., Boyd, J. H., Brown, A. P., Semmens, J. B. (2016). Limited privacy protection and poor sensitivity: is it time to move on from the Statistical Linkage Key-581? *Health information management journal 45*(2), 71-79.

This journal allows the author to use the published version of this manuscript within any book of which they are the author; evidence of this has been added below

16. **Randall, S. M.**, Brown, A. P., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2015). Privacy preserving record linkage using homomorphic encryption. *Proceedings of the First International Workshop on Population Informatics for Big Data, 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Sydney, Australia.

As mentioned in the footer of page 1 of the article, the copyright of this article remains with the authors.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Dec 03, 2016

This Agreement between Sean M Randall ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4001190301749
License date	Dec 03, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	International Journal of Cardiology
Licensed Content Title	Western Australia population trends in the incidence of acute myocardial infarction between 1993 and 2012
Licensed Content Author	Sean M. Randall,Renate Zilkens,Janine M. Duke,James H. Boyd
Licensed Content Date	1 November 2016
Licensed Content Volume Number	222
Licensed Content Issue Number	n/a
Licensed Content Pages	5
Start Page	678
End Page	682
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Enabling health research using administrative data: methodological improvements
Expected completion date	Jan 2017
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Requestor Location	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, WA 6102 Australia Attn: Sean M Randall
Total	0.00 AUD

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment

terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **T**ranslation : This permission is granted for non-exclusive world English rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any W**ebsite: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Pr eprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog

- ○ by updating a preprint in arXiv or RePEc with the accepted manuscript
- ○ via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
- ○ directly by providing copies to their students or to research collaborators for their personal use
- ○ for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement

– after the embargo period

- ○ via non-commercial hosting platforms such as their institutional repository
- ○ via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- – link to the formal publication via its DOI
- – bear a CC-BY-NC-ND license - this is easy to do
- – if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above:
Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.

Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.8

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

AMERICAN ACADEMY OF PEDIATRICS LICENSE TERMS AND CONDITIONS

Dec 02, 2016

This Agreement between Sean M Randall ("You") and American Academy of Pediatrics ("American Academy of Pediatrics") consists of your license details and the terms and conditions provided by American Academy of Pediatrics and Copyright Clearance Center.

License Number	4000710869378
License date	Dec 02, 2016
Licensed Content Publisher	American Academy of Pediatrics
Licensed Content Publication	Pediatrics
Licensed Content Title	Mortality After Burn Injury in Children: A 33-year Population-Based Study
Licensed Content Author	Janine M. Duke,Suzanne Rea,James H. Boyd,Sean M. Randall,Fiona M. Wood
Licensed Content Date	Apr 1, 2015
Licensed Content Volume Number	135
Licensed Content Issue Number	4
Licensed Content Pages	8
Type of Use	Dissertation/Thesis
Requestor type	Individual
Format	Print and Electronic
Portion	Full article
Order reference number	
Requestor Location	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, WA 6102 Australia Attn: Sean M Randall
Billing Type	Invoice
Billing Address	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, Australia 6102 Attn: Sean M Randall
Total	0.00 USD
Terms and Conditions	

AAP TERMS ANDCONDITIONS

The American Academy of Pediatrics grants permission to use the content cited above for the purpose stated. This letter shall serve as a receipt for payment of the permissions fee(s) and as an approval agreement.

1. The following credit line must appear:
   Reproduced with permission from Journal <Journal>, Vol. <Vol>, Page(s) <Pages>, Copyright © <Year> by the AAP

2. The requester guarantees to reprint the materials exactly as originally published. Obvious typographical errors maybe corrected. No deletions, alterations, or other changes may be made to the information or statistical data without the written consent of the American Academy of Pediatrics.
3. Rights granted herein are not exclusive and the American Academy of Pediatrics reserves the right to grant the same permission to others. Permission is granted for only the reproduction media specified.
4. Original artwork or copies of articles cannot be supplied, but PDF files may be downloaded from www.aappublications.org . Quantities of reprints and eprints can be obtained by contacting Terry Dennsteadt, Reprint Sales Manager – AAP Journals, The Walchli Tauber Group, Inc., 2225 Old Emmorton Road, Suite 201, Bel Air, MD 21046. 443.512.8899 x 112 office, 443.512.8909 fax, terry.dennsteadt@wt-group.com.
5. This permission is granted on a one-time, annual basis only. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given. Future use of this material is subject to the conditions stated herein. Gratis permissions are not issued for use in materials available for commercial sale, even for educational use.
6. If the permission fee for the requested use of our material is waived in this instance, please be aware future requests for AAP materials are subject to fees.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment. Provided that you have disclosed complete and accurate details of your proposed use, no license is effective unless and until full payment is received from you(either by publisher or by CCC) as provided in the CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted.
9. Warranties: Publisher makes no representations or warranties of any kind, express or implied, including but not limited to, accuracy, timeliness or completeness of the information contained in the licensed materials, or merchantability, title or fitness of a use for a particular purpose.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in writing signed by both requestor and publisher.
13. This permission, if permission has been granted for use of figures/tables/images, does not cover any third party copyrighted work which may appear in the material requested and does not apply to materials credited to publications other than American Academy of Pediatrics (AAP) journals. For materials credited to non-AAP journal publications, you will need to obtain permission from the publication referenced in the material legend or credit line before proceeding with usage of the materials. You agree to hold harmless and indemnify the AAP against any claims arising from your use of any content in your work that is credited to non-AAP sources.
14. This permission does not apply to and is not valid for photographs depicting identifiable individuals, including images where individuals' eyes have been blacked out or images depicting victims of abuse.
15. If the requester is translating the material, the following translation disclaimer must be included:
    The materials reused with permission from the American Academy of Pediatrics ("AAP") appeared originally in English, published by the AAP. The AAP assumes no responsibility for any inaccuracy or error in the contents of these materials, including any inaccuracy or error arising from the translation from English.
16. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Dec 02, 2016

This Agreement between Sean M Randall ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4000710647331
License date	Dec 02, 2016
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Australian and New Zealand Journal of Public Health
Licensed Content Title	Understanding the origins of record linkage errors and how they affect research outcomes
Licensed Content Author	James H. Boyd,Anna M. Ferrante,Katie Irvine,Michael Smith,Elizabeth Moore,Adrian Brown,Sean M. Randall
Licensed Content Date	Nov 20, 2016
Licensed Content Pages	1
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	Enabling health research using administrative data: methodological improvements
Expected completion date	Jan 2017
Expected size (number of pages)	200
Requestor Location	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, WA 6102 Australia Attn: Sean M Randall
Publisher Tax ID	EU826007151
Billing Type	Invoice
Billing Address	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, Australia 6102 Attn: Sean M Randall
Total	0.00 AUD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order**, is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY

QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.

- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS
Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.
The Creative Commons Attribution License
The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-
Creative Commons Attribution Non-Commercial License
The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License
The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)
Use by commercial "for-profit" organizations
Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.
Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Dec 02, 2016

This Agreement between Sean M Randall ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4000650895795
License date	Dec 02, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Computer Methods and Programs in Biomedicine
Licensed Content Title	Use of graph theory measures to identify errors in record linkage
Licensed Content Author	Sean M. Randall,James H. Boyd,Anna M. Ferrante,Jacqueline K. Bauer,James B. Semmens
Licensed Content Date	July 2014
Licensed Content Volume Number	115
Licensed Content Issue Number	2
Licensed Content Pages	9
Start Page	55
End Page	63
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Enabling health research using administrative data: methodological improvements
Expected completion date	Jan 2017
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Requestor Location	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, WA 6102 Australia Attn: Sean M Randall
Total	0.00 AUD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment

terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **T**ranslation : This permission is granted for non-exclusive world English rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any W**ebsite: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Pr eprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog

- by updating a preprint in arXiv or RePEc with the accepted manuscript
- via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
- directly by providing copies to their students or to research collaborators for their personal use
- for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement

– after the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

– link to the formal publication via its DOI
– bear a CC-BY-NC-ND license - this is easy to do
– if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above:
Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third

SPRINGER LICENSE
TERMS AND CONDITIONS

Dec 02, 2016

This Agreement between Sean M Randall ("You") and Springer ("Springer") consists of your license details and the terms and conditions provided by Springer and Copyright Clearance Center.

License Number	4000710206065
License date	Dec 02, 2016
Licensed Content Publisher	Springer
Licensed Content Publication	Springer eBook
Licensed Content Title	Application of Privacy-Preserving Techniques in Operational Record Linkage Centres
Licensed Content Author	James H. Boyd
Licensed Content Date	Jan 1, 2015
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	1
Author of this Springer article	Yes and you are a contributor of the new work
Order reference number	
Title of your thesis / dissertation	Enabling health research using administrative data: methodological improvements
Expected completion date	Jan 2017
Estimated size(pages)	200
Requestor Location	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, WA 6102 Australia Attn: Sean M Randall
Billing Type	Invoice
Billing Address	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, Australia 6102 Attn: Sean M Randall
Total	0.00 AUD
Terms and Conditions	

Introduction
The publisher for this copyrighted material is Springer. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).
Limited License
With reference to your request to reuse material on which Springer controls the copyright, permission is granted for the use indicated in your enquiry under the following conditions:

- Licenses are for one-time use only with a maximum distribution equal to the number stated in your request.
- Springer material represents original material which does not carry references to other sources. If the material in question appears with a credit to another source, this permission is not valid and authorization has to be obtained from the original copyright holder.
- This permission

• is non-exclusive
• is only valid if no personal rights, trademarks, or competitive products are infringed.
• explicitly excludes the right for derivatives.

- Springer does not supply original artwork or content.
- According to the format which you have selected, the following conditions apply accordingly:

• Print and Electronic: This License include use in electronic form provided it is password protected, on intranet, or CD-Rom/DVD or E-book/E-journal. It may not be republished in electronic open access.

• Print: This License excludes use in electronic form.

• Electronic: This License only pertains to use in electronic form provided it is password protected, on intranet, or CD-Rom/DVD or E-book/E-journal. It may not be republished in electronic open access.

For any electronic use not mentioned, please contact Springer at permissions.springer@spi-global.com.

- Although Springer controls the copyright to the material and is entitled to negotiate on rights, this license is only valid subject to courtesy information to the author (address is given in the article/chapter).
- If you are an STM Signatory or your work will be published by an STM Signatory and you are requesting to reuse figures/tables/illustrations or single text extracts, permission is granted according to STM Permissions Guidelines: http://www.stm-assoc.org/permissions-guidelines/

For any electronic use not mentioned in the Guidelines, please contact Springer at permissions.springer@spi-global.com. If you request to reuse more content than stipulated in the STM Permissions Guidelines, you will be charged a permission fee for the excess content.

Permission is valid upon payment of the fee as indicated in the licensing process. If permission is granted free of charge on this occasion, that does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

-If your request is for reuse in a Thesis, permission is granted free of charge under the following conditions:

This license is valid for one-time use only for the purpose of defending your thesis and with a maximum of 100 extra copies in paper. If the thesis is going to be published, permission needs to be reobtained.

- includes use in an electronic form, provided it is an author-created version of the thesis on his/her own website and his/her university's repository, including UMI (according to the definition on the Sherpa website: http://www.sherpa.ac.uk/romeo/);
- is subject to courtesy information to the co-author or corresponding author.

Geographic Rights: Scope

Licenses may be exercised anywhere in the world.

Altering/Modifying Material: Not Permitted

Figures, tables, and illustrations may be altered minimally to serve your work. You may not alter or modify text in any manner. Abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s).

Reservation of Rights

Springer reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction and (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

License Contingent on Payment

While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Springer or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received by the date due, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Springer reserves the right to take any and all action to protect its copyright in the materials.

Copyright Notice: Disclaimer

You must include the following copyright and permission notice in connection with any reproduction of the licensed material:

"Springer book/journal title, chapter/article title, volume, year of publication, page, name(s) of author(s), (original copyright notice as given in the publication in which the material was originally published) "With permission of Springer"

In case of use of a graph or illustration, the caption of the graph or illustration must be included, as it is indicated in the original publication.

Warranties: None

Springer makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

Indemnity

You hereby indemnify and agree to hold harmless Springer and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

No Transfer of License

This license is personal to you and may not be sublicensed, assigned, or transferred by you without Springer's written permission.

No Amendment Except in Writing

This license may not be amended except in a writing signed by both parties (or, in the case of Springer, by CCC on Springer's behalf).

Objection to Contrary Terms

Springer hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

Jurisdiction

All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in the Federal Republic of Germany, in accordance with German law.

Other conditions:

V 12AUG2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

ELSEVIER LICENSE TERMS AND CONDITIONS

Dec 02, 2016

This Agreement between Sean M Randall ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	3998560401993
License date	Nov 29, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of Biomedical Informatics
Licensed Content Title	Privacy-preserving record linkage on large real world datasets
Licensed Content Author	Sean M. Randall,Anna M. Ferrante,James H. Boyd,Jacqueline K. Bauer,James B. Semmens
Licensed Content Date	August 2014
Licensed Content Volume Number	50
Licensed Content Issue Number	n/a
Licensed Content Pages	8
Start Page	205
End Page	212
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Enabling health research using administrative data: methodological improvements
Expected completion date	Jan 2017
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Requestor Location	Sean M Randall Room 237 Building 400 Curtin University Kent St, Bentley Perth, WA 6102 Australia Attn: Sean M Randall
Total	0.00 AUD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions

established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those

established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **T**ranslation : This permission is granted for non-exclusive world English rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any W**ebsite: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Pr eprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group

    - directly by providing copies to their students or to research collaborators for their personal use
    - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- after the embargo period
    - via non-commercial hosting platforms such as their institutional repository
    - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.
Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.

Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.8

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Copyright Clearance Center RightsLink®

Home Account Info Help

Title:	Limited privacy protection and poor sensitivity:Is it time to move on from the statistical linkage key-581?
Author:	Sean M Randall, Anna M Ferrante, James H Boyd, Adrian P Brown, James B Semmens
Publication:	Health Information Management Journal
Publisher:	SAGE Publications
Date:	08/01/2016

Logged in as:
Sean Randall
Account #:
3001088594

LOGOUT

Gratis Reuse

- Without further permission, as the Author of the journal article you may:
  - post the accepted version (version 2) on your personal website, department's website or your institution's repository. You may NOT post the published version (version 3) on a website or in a repository without permission from SAGE.
  - post the accepted version (version 2) of the article in any repository other than those listed above 12 months after official publication of the article.
  - use the published version (version 3) for your own teaching needs or to supply on an individual basis to research colleagues, provided that such supply is not for commercial purposes.
  - use the accepted or published version (version 2 or 3) in a book written or edited by you. To republish the article in a book NOT written or edited by you, permissions must be cleared on the previous page under the option 'Republish in a Book/Journal' by the publisher, editor or author who is compiling the new work.
- When posting or re-using the article electronically, please link to the original article and cite the DOI.
- All other re-use of the published article should be referred to SAGE. Contact information can be found on the bottom of our 'Journal Permissions' page.

BACK CLOSE WINDOW