

**Centre for Population Health Research  
Office of Research and Development**

**Record Linkage Techniques: Exploring and developing data  
matching methods to create national record linkage infrastructure  
to support population level research**

**James Hutchison Boyd**

**This thesis is presented for the Degree  
of Doctor of Philosophy (Supplication)  
of  
Curtin University**

**December 2016**



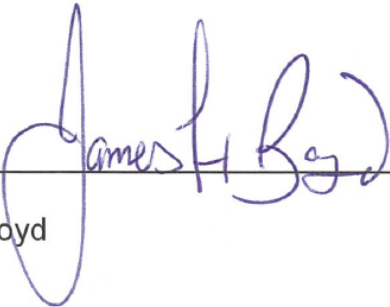
## Declaration

---

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

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

Signed:

  
James H Boyd

Date: 5<sup>th</sup> December 2016



## Thesis Summary

---

This thesis explores the development of national record linkage infrastructure in Australia using lessons learned from international examples, including the Scottish Record Linkage System. Such linkage capabilities enable the creation of very rich, longitudinal person-based research datasets of the health and healthcare experience of the whole population.

Over the last thirty years, research using record linkage has demonstrated its value in supporting policy and improving public services. The demand for linked data is strong, with linkage infrastructure being used internationally to service a range of linkage projects and/or programs of work; including government-based research and evaluation, healthcare utilisation, patient outcomes, resource allocation and epidemiological research. The development of responsive, agile and efficient infrastructure which can assess and manage privacy risks will enable research that can improve peoples' lives.

### Increasing data availability

Increasingly, 'life event' data is collected about an individual when they come in contact with organisations providing health, education and welfare services. This information includes data collections relating to hospital, school, police as well as birth, death and marriage. These collections are often mandated and covered under government legislation and curated through secure information governance procedures.

Australia is one of a small number of countries in the world with an accessible and wide-ranging healthcare system mainly funded through taxation and private healthcare arrangements. Responsibility for delivery of these health services is divided between the Australian Government (Commonwealth) and States (and Territories) and between public and private sector service providers. Data is generated using a standard set of data definitions and codes, to monitor these services, allowing the creation of comprehensive, high quality population-based health data collections.

Health data covers the broad range of medical services (both public and private sector), community health and statutory reporting. The healthcare information includes population-based data collections encompassing hospitalisation, emergency department attendance, births (including perinatal morbidity), cancer registrations, deaths, health screening services as well as primary care and community prescribing. These data collections are increasingly

supplemented by a range of specialised clinical and registry based data collections, covering stroke, cardiovascular disease, intensive care, renal disease and trauma.

## **Record Linkage**

Record linkage is a method used to integrate information from different sources to provide a complete picture of an individual's experience and service interaction over time. The combined datasets provide diverse and valuable resources for statistical and research projects without the time and cost associated with additional data collection. Record linkage systems have to handle large volumes of data and as a result, require complex organisational and technical infrastructure.

Developing and maintaining record linkage systems that integrate multiple datasets from different sources on a routine basis requires collaboration and data sharing across different organisations.

## **The development of National and International record linkage**

As linkage operations grow, there is a point where frequent manual linkage of the same datasets for a variety of different projects becomes inefficient and difficult to manage. With ever increasing computing power and data storage capacity, the early pioneers of record linkage for health research in both Canada and the UK (Oxford and Scotland) realised that there were opportunities for more permanent facilities. These linkage units recognised that enduring links created and routinely updated using 'enterprise' linkage infrastructure would provide a more efficient service.

The move to 'ongoing' linkage services has been adopted by almost all established linkage organisations and has seen the creation of large linked datasets available for research. These linkage repositories involve data from all government portfolios (not just health) and are benefiting from the recent developments in 'Big Data analytics'. Applying big data analytics allows researchers to gain valuable insights from a blend of structured, semi-structured and unstructured linked data.

In Australia, linked data is seen as a valuable and strategic resource and provides a cost effective way of using available data to support research and inform policy. Australia is a world leader in the development and operation of safe and secure data linkage systems, providing the research community access to linked data since the establishment of the WA Data Linkage System (WADLS) in 1995 and the Centre for Health Record Linkage (CHeReL) in 2006.

## National linkage in Australia

The Australian Government has made a substantial investment in building national record linkage infrastructure for Australia through the establishment of the Population Health Research Network (PHRN) funded by the National Collaborative Research Infrastructure Strategy (NCRIS), Education Investment Fund-Super Science Initiative (EIF-SSI) and the Collaborative Research Infrastructure Strategy (CRIS). The principal purpose of the PHRN is to build a nationwide record linkage infrastructure capable of securely and safely linking data collections from a wide range of sources including within and between jurisdictions and across sectors. The resulting infrastructure provides Australia with a unique international record linkage platform that will significantly increase research capacity.

This thesis describes the challenges of building large scale record linkage infrastructure using the lessons learned from other national and international systems. The move to the flexible routine linkage of 'Big Data' comes with challenges relating to data processing (data standards, data management, scalability and quality) and governance (information frameworks, confidentiality and privacy protection). The design includes solutions to some of the issues from existing systems and develops new and innovative techniques to address problems, not previously solved in large routine systems.

## Research translation

Finally, the benefits of record linkage are demonstrated through application in a series of epidemiological studies. The infrastructure is first tested through a proof of concept project which validates linkage and epidemiological benefits of cross-jurisdictional linkage in Australia exploring mortality following hospitalisation. Both the advantages and boundaries of the process are discussed as patient pathways are followed across four state borders (Western Australia, South Australia, New South Wales and Queensland). The project demonstrates the power of national linkage and the advantages in terms of power, completeness and accuracy of study populations.

An international collaborative study outlines the possibility of pulling together study populations from other countries with compatible linkage infrastructure and data collections. The Burns study shows the research potential by combining data from both Scotland and Western Australia to increase statistical power that has resulted in translation changes to clinical practice and the development of clinical guidelines. This unique analysis combined hospital records and cancer registration data from both systems at a unit level. It allowed statistical analysis of the data demonstrating increased risk of cancer by type within

subgroups. This is an important finding which will help inform follow-up services and screening for burns patients.

This study formed part of the Western Australian Population-based Burn Injury Project – a retrospective cohort investigation - using Western Australian population-based linked data for burn and uninjured groups. This project has identified increased long term all-cause mortality for paediatric and adult burn trauma patients. Our research of post burn morbidity has identified increased cardiovascular, musculoskeletal, respiratory and gastrointestinal morbidity in terms of post burn hospital admissions. Findings of increased mortality and morbidity after both severe and minor burns are significant since the majority of burn injury admissions in Western Australia, as for other developed countries, are for minor burn injuries. In addition, recent evidence generated by ‘basic scientific’ research at the Burn Injury Research Unit (BIRU) strongly implicates immune changes after burn injury leading to increased susceptibility to infection, cancer, bone loss and cardiac changes.

The data generated from this project has filled gaps in current data and provided the evidence base required to inform best clinical practice and to direct strategic health policy in the treatment and long-term management of burn injury to improve the health, wellbeing and ‘quality of life’ of burn injury patients.

Translation of outcomes to affect health policy has been driven by Professor Fiona Wood, the Director of the Burn Service of Western Australia (BSWA), and her research team. Morbidity and economic outcomes have been presented to the Chief Medical Officer (Dr Gary Geelhoed), Executive Director of Public Health (Dr Tarun Weeramanthri), and the Injury and Trauma Clinical Network, Department of Health Western Australia. The objectives of this network are to establish state-based strategic directions, influence policy development, assist in resource allocation, and develop preventative strategies. The BSWA has responsibilities for the ongoing development of both the WA and National ‘Burn Injury Model of Care’ as part of the WA and National Health Reform Process. The costs of long-term hospital service use attributable to burns will be used to inform health policy in regards to resource allocation as well as burn prevention strategies to reduce burn injury and subsequent post burn morbidity and hospital costs.

## **Conclusion**

Record linkage provides a cost effective alternative for research project as it minimises expensive and intrusive data collection. Although the science is proven and well established, advances in computing power continue to allow developments in the size and complexity of



linkage projects. As research exploits data from different settings, record linkage systems have to adapt to allow multi-faceted large linkage projects without limitations in terms of time, cost and resources. This thesis demonstrates the value of national linkage and how this has been achieved for the Australian research community.



## Acknowledgments

---

Having spent many years developing record linkage services for government and university researchers in Scotland, Curtin University and the Centre for Population Health Research provided me with a fantastic opportunity to expand research around record linkage methods, techniques and applications in Australia.

I would like to thank Professor James Semmens whose enthusiasm and support made the decision to move to Australia easy and the time spent on the research project a pleasure. I am grateful for your guidance, support, feedback and most importantly your enthusiasm which is always inspiring!

Thank you to Associate Professor Anna Ferrante for your advice and support. I have benefitted greatly from the knowledge you have shared and still marvel at how much we sorted out in the first month of the Centre for Data Linkage.

The Centre for Data Linkage has been a fantastic place to work. I thank you all for your dedication, support and very high standards and look forward to many more years of productive research together.

And finally to Margo, my long suffering partner who moved across the world to support me; I could not have done it without you!

Slàinte mhath!



# Centre for Population Health Research

## Director's recommendation

---

### Thesis content for assessment

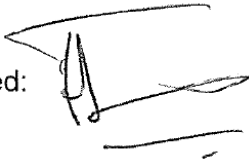
The National Collaborative Research Strategy (NCRIS) was initiated in Australia in 2009 and supported the development of national data linkage infrastructure as a priority theme within the Population Health Research Network (PHRN). Within this initiative, the Centre for Data Linkage (CDL) was established to provide cross jurisdictional national data linkage among the States and Territories of Australia. To meet the needs of this strategic initiative, Associate Professor James Boyd was specifically targeted for employment by Curtin University through the Centre for Population Health Research for his internationally recognised expertise in data linkage. Associate Professor Boyd was working with the Scottish Government and had played a leading role in the development of data linkage capacity in health and education. He accepted Curtin's invitation and commenced work at Curtin University in April 2009. Given these circumstances, I suggest his 'employment at Curtin should be considered to be of continuous service' from his previous employment through the Scottish Government.

Thus, while his thesis by supplication is based on 11 published papers, one book chapter and a published letter, I have requested he also include five significant papers published during his tenure with the NHS in Scotland as 'Supporting Publications'. While his full list of publications accounts for more than 50 publications, the significance of these five supporting publications demonstrate his expertise and contribution to data linkage and support the basis for Curtin efforts to strategically 'poach' him from Scotland. His Scottish experience is directly relevant to his task to establish the Centre for Data Linkage at Curtin University as part of the NCRIS-PHRN initiative.

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. It does not diminish the role of the primary author (Papers 1-5, 13 (published letter) and 14 (book chapter)). Associate Professor Boyd has also provided a leading role in the use of data linkage platforms for health research as evidenced by his collaborative role in papers 6-8. The final three papers included (papers 9-11) outline important developments linkage quality and the development of Privacy Preserving Record Linkage, as a novel way to address important legal and administrative concerns for the linkage of specific types of sensitive datasets. Associate Professor Boyd is part of the leadership team supporting this

initiative, which has recently expanded to include collaboration from groups in the United Kingdom (London, Scotland, Wales), Germany, Canada and Australia (NSW, WA).

Signed:

A handwritten signature in black ink, appearing to be 'J. Semmens', written over a horizontal line.

Date: 5.12.16

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

## Research output from this thesis

---

This thesis contains the following peer-reviewed scientific publications. These papers have been published in quality medical and public health journals.

### Published manuscript(s)

1. **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 services research 12.1 (2012): 480.
2. **Boyd JH**, Randall SM, Ferrante, AM Bauer JK, Brown AP, Semmens JB. *Technical challenges of providing record linkage services for research* (2014) BMC Medical Informatics and Decision Making, 14 (1), art. no. 23.
3. **Boyd JH**, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, Dickinson T. *A simple sampling method for estimating the accuracy of large scale record linkage projects.* Methods of information in medicine. 2016;55(3):276-83.
4. **Boyd JH**, Randall SM, Ferrante AM, Bauer JK, McInnery K, Brown AP, Spilsbury K, Gillies M and Semmens JB. *Accuracy and completeness of patient pathways - the benefits of national data linkage in Australia.* BMC health services research 15.1 (2015): 312.
5. **Boyd JH**, Wood FM, Randall SM, Fear MW, Rea S, Duke, JM. (2016). Effects of pediatric burns on gastrointestinal diseases: A population-based study. The Journal of Burn Care & Research (2016). doi: 10.1097/BCR.0000000000000415
6. Spilsbury K, Rosman D, Alan J, **Boyd JH**, Ferrante A, Semmens JB. *Cross border hospital use: An analysis using data linkage across four Australian states.* Medical Journal of Australia 202.11 (2015): 582-586.
7. Ferrante AM and **Boyd JH** (2012) A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics (45)165-172.
8. Duke J M, Bauer J, Fear MW, Rea S, Wood FM, **Boyd J** (2014). Burn injury, gender and cancer risk: population-based cohort study using data from Scotland and Western Australia. BMJ open, 4(1), e003845.4
9. Randall SM, Ferrante AM, **Boyd JH**, Bauer JK, Semmens JB. *Privacy-preserving record linkage on large real world datasets.* Journal of Biomedical Informatics 2014; DOI: 10.1016/j.jbi.2013.12.003.
10. Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. *The effect of data cleaning on record linkage quality.* BMC Medical Informatics and Decision Making 2013; 13 (64): e1-e10.
11. Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. *Use of graph theory measures to identify errors in record linkage.* Computer Methods and Programs in Biomedicine. Volume 115, Issue 2, July 2014, Pages 55-63.

## Book Chapter(s)

12. **Boyd JH**, Randall SM, Ferrante AM. *Application of privacy preserving techniques in operational record linkage centres*. Medical Data Privacy Handbook. Springer International Publishing, 2015. 267-287.

## Published letter(s)

13. **Boyd JH**, Ferrante AM, Irvine K, Smith M, Moore E, Brown AP, Randall SM. *Understanding the Origins of record linkage error and how they affect research outcomes*. Australia and New Zealand Journal of Public Health doi: 10.1111/1753-6405.12597.

## Supporting manuscripts (1996 – 2001)

14. Walsh D, Smalls M, **Boyd J**. *Electronic health summaries--building on the foundation of Scottish Record Linkage system*.(2001) Medinfo, 10 (Pt 2), pp. 1212-1216.
15. Capewell S, Kendrick S, **Boyd J**, Cohen G, Juszczak E, Clarke J. *Measuring outcomes: One month survival after acute myocardial infarction in Scotland*. (1996) Heart, 76 (1), pp. 70-75.
16. Capewell S, MacIntyre K, Stewart S, Chalmers JWT, **Boyd J**, Finlayson A, Redpath A, Pell JP, McMurray JJV. *Age, sex, and social trends in out-of-hospital cardiac deaths in Scotland 1986-95: A retrospective cohort study* (2001) Lancet, 358 (9289), pp. 1213-1217.
17. MacIntyre K, Capewell S, Stewart S, Chalmers JWT, **Boyd J**, Finlayson A, Redpath A, Pell JP, McMurray JJV. *Evidence of improving prognosis in heart failure: Trends in case fatality in 66 547 patients hospitalized between 1986 and 1995* (2000) Circulation, 102 (10), pp. 1126-1131.
18. MacIntyre K, Stewart S, Capewell S, Chalmers JWT, Pell JP, **Boyd J**, Finlayson A, Redpath A, Gilmour H, McMurray JJV. *Gender and survival: A population-based study of 201,114 men and women following a first acute myocardial infarction* (2001) Journal of the American College of Cardiology, 38 (3), pp. 729-735.

## International Conference presentation(s)

1. **Boyd JH**, Ferrante AM, Randall S, Bray J, O'Shea, A, Semmens JB. Development of National Data Linkage: A linkage system for the 21st Century. 2012 International Health Data Linkage Network Conference (IHDLN). Perth, Western Australia, May 2012
2. **Boyd JH**, Ferrante AM, Randall SM, Bauer, J, Gillies M, Semmens JB. Developing National Data Linkage Infrastructure in Australia: SHIP International Conference: Exploiting existing data for health research. St Andrews, Scotland, September 2013.
3. Randall SM, **Boyd JH**, Ferrante AM, Bauer J, Gillies M, Semmens JB. Privacy Preserving record linkage on large real world datasets. Conference: International Health Data Linkage Conference, Vancouver, Canada, April 2014.
4. Ferrante AM, **Boyd JH**, Randall SM, Brown AP, Semmens JB. How do you measure up? Methods to assess linkage quality. Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.



5. Randall SM, Ferrante AM, Brown AP, **Boyd JH**, Semmens JB. Assessing the impact of different grouping methods: time to rethink and regroup? Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.
6. **Boyd JH**. Using record linkage to examine long-term effects of burn injury: The Western Australian Population-based Burn Injury Project. Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

*I warrant that I have obtained, where necessary, permission from the copyright owners to use any third party copyright material reproduced in the thesis (e.g. questionnaires, artwork, unpublished letters), or to use any of my own published work (e.g. journal articles) in which the copyright is held by another party (e.g. publisher, co-author).*

## Competitive grants

- 2016: **Boyd JH**, Ferrante AM, Brown AP, Randall SM, Semmens JB Australia-Germany Joint Research Co-operation Scheme. DAAD 2016 (Funding commencing in 2017). Australia Universities. (\$12,220)
- Boyd JH**, Ferrante AM, Brown AP, Semmens JB. National Collaborative Research Infrastructure Strategy (NCRIS 2016). Population Health Research Network. Department of Education and Training. (\$485,465)
- Boyd JH**, Ferrante AM, Semmens JB. National Collaborative Research Infrastructure Strategy (NCRIS 2015). Population Health Research Network. Department of Education and Training. (\$479,880)
- Boyd JH**, Ferrante AM, Semmens JB. National Collaborative Research Infrastructure Strategy (NCRIS 2013). Population Health Research Network. Department of Education and Training. (\$100,000)
- 2015: Semmens JB, **Boyd JH**, Ferrante AM. Education Investment Fund Super Science Initiative (EIF-SSI Additional Funding). Population Health Research Network. Department of Education and Training. (\$45,334)
- Boyd JH**, Ferrante AM, Semmens JB. National Collaborative Research Infrastructure Strategy (NCRIS 2013). Population Health Research Network. Department of Innovation, Industry, Science & Research (DIISR). (\$539,999)
- 2014: Semmens JB, **Boyd JH**, Ferrante AM. Education Investment Fund Super Science Initiative (EIF-SSI). Population Health Research Network. Department of Education and Training. (\$55,833)
- Boyd JH**, Ferrante AM. National Collaborative Research Infrastructure Strategy (NCRIS). Population Health Research Network Data Delivery System (DDS) - Phase 3 Development. Department of Education and Training. (\$10,000)
- Boyd JH**, Ferrante AM, Semmens JB. Collaborative Research Infrastructure Scheme (CRIS): Population Health Research Network. Department of Innovation, Industry, Science & Research (DIISR). (\$544,500)
- 2013: Semmens JB, **Boyd JH**, Ferrante AM. Education Investment Fund Super Science Initiative (EIF-SSI) Population Health Research Network. Department of Education and Training. (\$279,159)
- Boyd JH**, Ferrante AM. National Collaborative Research Infrastructure Strategy (NCRIS). Population Health Research Network. Data Delivery System (DDS) - Phase 3 Development. Department of Education and Training. (\$170,000)
- 2012: Semmens JB, **Boyd JH**, Ferrante AM. National Collaborative Research Infrastructure Strategy (NCRIS). Population Health Research Network. Department of Innovation, Industry, Science & Research (DIISR). (\$124,275)
- Semmens JB, **Boyd JH**, Ferrante AM. Education Investment Fund Super Science Initiative (EIF-SSI). Population Health Research Network. Department of Innovation, Industry, Science & Research (DIISR). (\$837,473)

- Boyd JH**, Ferrante AM. National Collaborative Research Infrastructure Strategy (NCRIS). Population Health Research Network. Data Delivery System (DDS) - Phase 2 Development Department of Innovation, Industry, Science & Research (DIISR).(\$285,000).
- 2011: Semmens JB, **Boyd JH**, Ferrante AM. National Collaborative Research Infrastructure Strategy (NCRIS). Population Health Research Network. Department of Innovation, Industry, Science & Research (DIISR). (\$1,118,475)
- 2010: Semmens JB, **Boyd JH**, Ferrante AM. National Collaborative Research Infrastructure Strategy (NCRIS). Population Health Research Network. Department of Innovation, Industry, Science & Research (DIISR). (\$1,134,000)



## Table of Contents

---

<b>Declaration</b> .....	<b>iii</b>
<b>Thesis Summary</b> .....	<b>v</b>
Increasing data availability .....	v
Record Linkage.....	vi
The development of National and International record linkage.....	vi
National linkage in Australia .....	vii
Research translation .....	vii
Conclusion.....	viii
<b>Acknowledgments</b> .....	<b>xi</b>
<b>Centre for Population Health Research</b> .....	<b>xiii</b>
<b>Director’s recommendation</b> .....	<b>xiii</b>
Thesis content for assessment.....	xiii
<b>Research output from this thesis</b> .....	<b>xv</b>
Published manuscript(s).....	xv
Book Chapter(s).....	xvi
Published letter(s).....	xvi
Supporting manuscripts (1996 – 2001) .....	xvi
International Conference presentation(s) .....	xvi
Competitive grants .....	xviii
<b>List of Figures</b> .....	<b>5</b>
<b>Abbreviations and Acronyms</b> .....	<b>7</b>
<b>Glossary</b> .....	<b>9</b>
<b>Exegesis</b> .....	<b>15</b>
Thesis Abstract .....	<b>15</b>
Background .....	<b>15</b>
Record Linkage Methods .....	<b>15</b>
Study Aims.....	<b>16</b>
Results .....	<b>16</b>
Conclusion .....	<b>17</b>
Aims and Objectives .....	<b>19</b>
Thesis Overview .....	<b>22</b>
<b>Chapter 1</b> .....	<b>27</b>
1.1. Introduction .....	<b>29</b>
1.2. Record Linkage Methodology.....	<b>29</b>

1.3. Components of the data linkage process .....	31
1.4. National and International record linkage developments .....	34
1.4.1. The Oxford Record Linkage Study (ORLS) .....	34
1.4.2. The Manitoba Population Health Information Systems .....	34
1.4.3. Population Data British Columbia.....	35
1.4.4. Secure Anonymised Information Linkage (SAIL).....	35
1.4.5. The Western Australia Data Linkage System (WADLS) .....	36
1.4.6. The Centre for Health Record Linkage (CHeReL) .....	37
1.5. The first routine national record linkage system .....	39
1.5.1. Record Linkage in Scotland .....	39
1.5.2. Early development in Scotland.....	39
1.5.3. Year 2000 (Y2K) redevelopment.....	40
1.5.4. 'One-pass' Linkage and the Best Link Principle .....	41
1.5.5. Community Health Index (CHI) number .....	41
1.5.6. Linking health and social data in Scotland .....	42
1.5.7. Research and development around record linkage .....	42
1.6. Supporting Manuscript .....	45
<b>Chapter 2 .....</b>	<b>53</b>
2.1. Data Linkage development in Australia .....	55
2.2. Population Health Research Network.....	55
2.3. Centre for Data Linkage .....	56
2.4. Designing Secure Linkage Infrastructure .....	56
2.5. Cross-Jurisdictional Operational Model.....	57
2.6. Designing National linkage infrastructure .....	60
2.7. Functional requirements.....	60
2.8. Development of an automated production linkage system .....	61
2.9. Performance evaluation of Linkage Engines .....	63
2.10. Linkage management and matching engine.....	63
2.11. Governance .....	64
2.12. Ethics approval .....	64
2.13. Conclusion .....	64
2.14. Published Manuscript(s). .....	67
<b>Chapter 3 .....</b>	<b>99</b>
3.1. Privacy challenges in Record Linkage.....	101
3.2. Data governance.....	101
3.3. Operating model and data flows.....	102
3.3.1. Centralised Model .....	102
3.3.2. Separated Model, with centralised clinical data.....	103
3.3.3. Separated Model, with no centralised data repository .....	105
3.3.4. Operational Models involving Multiple Linkage Units.....	106

3.4. Privacy Preserving Record Linkage Methods.....	106
3.5. Balancing the privacy constraints.....	108
3.6. Application of Privacy Preserving Record Linkage .....	110
3.7. Conclusion.....	112
3.8. Book Chapter.....	115
3.9. Published Manuscript.....	139
<b>Chapter 4.....</b>	<b>149</b>
4.1. Origins of record linkage error.....	151
4.2. Linkage quality.....	151
4.2.1. Metrics for measuring errors in linkage .....	152
4.2.2. Estimating linkage quality.....	153
4.3. Techniques to optimise quality and reduce errors .....	155
4.3.1. Quality and quantity of incoming data .....	155
4.3.2. Data cleaning (and standardisation).....	156
4.3.3. Clerical assessment and review.....	158
4.3.4. Identifying errors to improve linkage quality .....	158
4.3.5. One-off assessments of linkage quality.....	159
4.3.6. Sensitivity analysis.....	159
4.3.7. Graph theory – Picturing the problem.....	159
4.3.8. Automated quality tools.....	160
4.3.9. Rule-based clerical intervention .....	161
4.4. Impact of linkage quality on research outcomes.....	161
4.5. Mitigating the impact of linkage error.....	162
4.6. Conclusion.....	163
4.7. Published Manuscript(s).....	165
4.8. Published Letter.....	199
<b>Chapter 5.....</b>	<b>203</b>
5.1. Proof of Concept – Research using linked data .....	205
5.2. Indicators of hospital mortality.....	205
5.3. Project phases .....	205
5.4. Data linkage.....	206
5.4.1. Data summary .....	206
5.4.2. Linkage strategy .....	207
5.4.3. Cross-jurisdictional data linkage results .....	207
5.4.4. Blocking efficiency .....	211
5.4.5. Comparison of CDL Linkages with Jurisdictional Linkages.....	212
5.4.6. Linkage accuracy.....	213
5.5. Epidemiology .....	215
5.6. Conclusion.....	215

5.7. Published Manuscript(s).....	217
<b>Chapter 6</b> .....	<b>235</b>
6.1. Data Availability and Use .....	237
6.2. Research Translation Case Studies.....	237
6.3. Research using Linked Data in Scotland.....	238
6.3.1. Studies Linking ISD Data for Epidemiology (SLiDE).....	240
6.3.2. SLiDE Research Summary .....	244
6.4. Burn injury research in Western Australia .....	245
6.4.1. Burn injury and cancer risk in Western Australia and Scotland .....	245
6.4.2. Western Australian Population-based Burn Injury Project .....	246
6.4.3. Effects of paediatric burn injury on gastrointestinal diseases .....	248
6.5. Conclusion.....	249
6.6. Published Manuscript(s).....	251
6.7. Supporting Manuscript(s) .....	273
<b>Chapter 7</b> .....	<b>307</b>
7.1. Conclusion .....	309
7.2. Solid infrastructure .....	309
7.3. A career in record linkage .....	310
7.4. An unexpected journey .....	310
7.5. A vision for the future .....	312
7.6. Professional Report(s) .....	315
<b>References</b> .....	<b>339</b>
<b>Appendix 1</b> .....	<b>347</b>
Conference Presentation Abstracts.....	347
Development of the National Linkage System: a linkage system for the 21st Century	349
Developing National Data Linkage Infrastructure in Australia .....	350
Privacy preserving record linkage on large real world datasets .....	351
How do you measure up? Methods to assess linkage quality.....	352
Assessing the impact of different grouping methods: time to rethink and regroup? ....	354
Using record linkage to examine long-term effects of burn injury: The Western Australian Population-based Burn Injury Project .....	355
<b>Appendix 2</b> .....	<b>357</b>
Statements of contribution.....	357
<b>Appendix 3</b> .....	<b>375</b>
Copyright statements .....	375



## List of Figures

---

Figure 1: Cross-jurisdictional data flows .....	60
Figure 2: Linkage processes .....	63
Figure 3: A centralised model: Data providers give their full datasets to the linkage unit, who link and then pass on the content data required for research to the researcher.....	104
Figure 4: The data provider splits their data, sending personal identifiers to the linkage units, and clinical content to the client services team. The linkage unit then provides the linkage map to the client services team who join it to content data to create datasets for research and analysis. ....	105
Figure 5: In the absence of a repository of clinical data, this is supplied to the researcher by the data provider .....	106
Figure 6: Privacy preserving record linkage process .....	108
Figure 7: Privacy Preserving Constraints .....	109
Figure 8: Creating a Bloom filter: a simple example using bigrams .....	112
Figure 9: Privacy Preserving Constraints .....	154
Figure 10: Graph theory metrics.....	161
Figure 11: Datasets provided to CDL .....	208
Figure 12: Number of pairs found in each linkage .....	209
Figure 13: Nature of data linkage of WA, NSW, SA and QLD data.....	210



## Abbreviations and Acronyms

Abbreviation/Acronym	Definition
ACE	Angiotensin Converting Enzyme
ACSQHC	Australian Commission on Safety and Quality in Health Care
ACT	Australian Capital Territory
AMI	Acute Myocardial Infarction
AR%	Attributable Risk Percent
ARIA	Accessibility Remoteness Index of Australia
BCLHD	British Columbia Linked Health Database
BIRU	Burn Injury Research Unit
BSWA	Burn Service of Western Australia
CDL	Centre for Data Linkage
CHD	Coronary Heart Disease
CHI	Community Health Index
CHeReL	Centre for Health Record Linkage
CI	Confidence Interval
CRIS	Collaborative Research Infrastructure Strategy
CUPLE	CUSTOMisable Probabilistic Linkage Engine
DDS	Data Delivery System
DMS	Data Management System
DoHWA	Department of Health, Western Australia
EDDC	Emergency Department Data Collection
EIF	Education Investment Fund
ETL	Extract, Transfer and Load
FP	False Positive
FN	False Negative
GP	General Practice
HACC	Home and Community Care Program
HIRU	Health Information Research Unit (Swansea University)
HREC	Human Research Ethics Committee
HMDS	Hospital Morbidity Data System
HR	Hazard Ratio
HSMR	Hospital Standardised Mortality Ratio
IaaS	Infrastructure as a Service
ICD	International Classification of Diseases
ICT	Information and Communication Technology
IRR	Incidence Rate Ratio
ISD	Information Services Scotland
MCHP	Manitoba Centre for Health Policy
MLK	Master Linkage Key
MPI	Master Patient Index
MRR	Mortality Rate Ratio
NCRIS	National Collaborative Research Infrastructure Strategy

Abbreviation/Acronym	Definition
NHS	National Health Service
NHSAR	NHS Administrative Register
NLS	National Linkage System
NSW	New South Wales
NT	Northern Territory
OR	Odds Ratio
ORLS	Oxford Record Linkage Study
PHRN	Population Health Research Network
PoC	Proof of Concept
PopData	Population Data British Columbia
PPRL	Privacy Preserving Record Linkage
PY	Person Years
QLD	Queensland
RLG	Research Linkage Group
SA	South Australia
SAAP	Supported Accommodation Assistance Program
SAIL	Secure Anonymised Information Linkage Databank
SCI	Scottish Care Information
SEIFA	Socio-Economic Indexes of Areas
SLIDE	Studies Linking ISD Data for Epidemiology
SLK	Statistical Linkage Key
SMC	Secure Multi-party Computation
SMR	Scottish Morbidity Record
SMR01	Scottish Morbidity Record collection (Scheme 1)
SMR04	Scottish Morbidity Record collection (Scheme 4)
SSI	Education Investment Fund Super Science Initiative
SURE	Secure Unified Research Environment
TBSA	Total Body Surface Area
TKI	Telethon Kids Institute
TP	True Positive
TN	True Negative
UK	United Kingdom
UPI	Unique Patient Identifier
US	United States
UWA	University of Western Australia
WA	Western Australia
WADLB	Western Australian Data Linkage Branch
WADLS	Western Australian Data Linkage System
Y2K	Year 2000

## Glossary

---

Term	Definition
Ad hoc data linkage	This involves the linkage of two or more datasets for a specific purpose and a specific often non-ongoing project, using a specific set of input datasets. Ad hoc data linkage does not involve the maintenance of a master linkage file and master linkage key.
Administrative data	Information that is collected for the purpose of, or in the process of, service delivery; such as providing health care (National Hospital Morbidity Database), responding to the legal requirements of registering particular events (births and deaths registration data) or providing a particular service.
Algorithm	A process or set of rules used for calculation or problem solving.
Blocking	In data linkage, blocking reduces the number of comparisons needed by only comparing record pairs where links are more likely to be found. Records on each file are placed into blocks so that only record pairs that agree on certain data items are compared.
Blocking variables	Variables used in partitioning records into blocks. Only records having the same value in a blocking variable are compared. Blocking variables must be stable, accurate and available on all files to be linked. Examples of blocking variables are first and last name, components of first and last name, sex, components of date of birth (e.g. month of birth or year of birth) and components of usual place of residence.
Clerical review	A manual review of record pairs whose link status cannot be automatically determined from their linkage weights or linkage probabilities. Clerical review helps determine the link status of these record pairs. Clerical review can also be used to obtain a quality assessment of a linkage.
Clerical assessment	A manual review of the validity or accuracy of the link status assigned to record pairs. The result of this assessment will assess whether the linked record pairs are true links or false links, true non-links or false non-links.
Confidentiality	Treatment of information about an individual or entity in a manner that will not disclose the identity of that individual or entity.

Term	Definition
Comparison record pair	Any pair of records being compared to determine whether or not they belong to the same person or entity.
Content data	De-identified service or administrative data collected by agencies and used by researchers.
Coverage	The extent to which a dataset captures the population in scope.
Data cleaning	The process of editing data to remove errors such as illogical and out-of-scope values, and data entry errors, such as typographical errors and transposed values. In data linkage, data cleaning may also encompass data standardisation. See also data standardisation.
Data standardisation	The process of making different datasets comparable and compatible, and conform to the same quality rules, in terms of structure of dataset, scope, completeness, coding, structure and spelling of variable names, and range and format of data values.
Data custodian	The authority, body or person responsible for the safe custody, transport and storage of data, and implementation of business rules regarding use of the data. Data custodians may either have collected the data themselves or they may have legal and administrative custody of it on behalf of the owner or collector of the data.
Data integration	The process of merging together content data using linkage keys.
Data linkage	The process of bringing together information belonging to the same person, event or place, into a single record of information. See Record Linkage.
Data separation	The process of dividing or separating data into two components - Demographic data and Content data. See also Separation process.
De-identification	Processes for removing identifying information from datasets, most commonly to protect the privacy of individuals.
De-identified data	Data which does not contain personal information or from which the identity of the individual to whom it pertains cannot be reasonably ascertained.

Term	Definition
Demographic data	Variables that are common to the data files being linked, and are used for matching records in the data linkage process. Examples of linking variables include data items of personal information: first name(s), last name, sex, date of birth, usual place of residence and country of birth. See also linking variables.
Deterministic linkage	Deterministic linkage ranges from simple joining of two or more datasets by a reliable and stable key to sophisticated stepwise algorithmic linkage. See simple deterministic linkage and stepwise deterministic linkage.
Dynamic data linkage system	A system of data linkage that involves the ongoing linkage of core datasets and the permanent maintenance of a master linkage file and master linkage key.
False-negative link	A pair of records belonging to the same individual or entity that is incorrectly assigned as non-matches or as not belonging to the same individual or entity.
False-negative rate	The proportion of all record pairs belonging to the same individuals or entities that are incorrectly assigned as non-links.
False-positive link	A pair of records belonging to two different individuals or entities that are incorrectly assigned as links.
False-positive rate	The proportion of all record pairs belonging to two different individuals or entities that are incorrectly assigned as links.
Identified data	Data that allow the identification of an individual, either directly or indirectly (potentially identifiable), are referred to as "identified data". Such data are deemed to be confidential.
Link	A decision that two records correspond to the same person or entity.
Linked	The status of a record that has passed through the data linkage process and was linked to a record from the other file.
Linking variables	Variables that are common to the data files being linked, and are used for comparing records. Examples of linking variables include first name(s), last name, sex, date of birth, usual place of residence and country of birth. Some linking variables can also be used as blocking variables. See also blocking variables and matching variables.

Term	Definition
Linkage Key	The codes created and stored by a data linkage unit that can be used to group records that refer to the same person or entity.
Linkage Map	A file of Linkage Keys.
Master Linkage Key (MLK)	The codes created and stored by a data linkage unit that can be used to group records that refer to the same person or entity.
Match	A record pair containing information that relates to the same unit. See also Link, Non-link, Non-match.
Match accuracy rate	Proportion of all record pair comparisons that are true positives (TP) or true negatives (TN). The denominator for this rate is the number of all record pair comparisons, while the numerator is the number of record pairs that are correctly classified as true matches or false matches.
Matching variables	See Linking variables.
MIDSPAN	MIDSPAN is the name used for the large occupational and general population health surveys, based in the West of Scotland, which began in the 1960s and involved nearly 30,000 people.
Non-link	A decision that two records do not correspond to the same person or entity.
Non-match	A record pair that contains information that relates to different people or entities.
Precision or positive predictive value	The proportion of all classified links that are true links as opposed to classified links that are false links. It is calculated by dividing the number of links that are ascertained as true, by the total number of classified links.
Privacy	The right of a person or group of people to keep their lives and personal affairs out of public view, and to control the flow of information about themselves.
Probabilistic linkage	A method of record linkage that uses the probabilities of agreement and disagreement between a range of linkage variables.
Recall	Recall (also known as sensitivity) is the proportion of true links that are identified through the matching process. It is calculated by dividing the number of links that are ascertained as true, by the total number of true links.



Term	Definition
Record linkage	The process of bringing together two or more sets of information belonging to the same person, event or place, into a single record of information, in a way that protects individual privacy. See Data linkage.
Record pairs	See comparison record pairs.
Sensitivity or true-positive rate	The proportion of all records in a file or database with a match in another file that were correctly accepted as a link.
Separation principle	A best practice model where roles, functions and data are clearly delineated. Personnel involved in the project will only have access to the selected data that is required for the particular operation they are undertaking. For example, staff undertaking data linkages will only access identifying variables (such as names and dates of birth), while staff undertaking merging will only access content (de-identified) variables. Refer to Kelman, Bass and Holman (2002).
Separation process	The process of dividing or separating data into two components - Demographic data and Content data. See also Separation process.
SMR01	Scottish Morbidity Record Scheme (Scheme 1) – Non Obstetric, Non Psychiatric inpatient and daycase discharges.
SMR04	Scottish Morbidity Record Scheme (Scheme 4) – Psychiatric inpatient admissions.
Specificity or true-negative rate	The proportion of all records on one file or database that have no match in the other file that were correctly not accepted as a link.
Statistical linkage key (SLK)	A code used in data linkage that replaces a person's identifiable data to protect the person's identity. It is generated from elements of an individual's personal demographic data and attached to de-identified data relating to the services received by that individual.
True-positive link	Two records that truly do correspond to the same person or entity. See Link, Non-link, True non-match.
Unlinked	The status of a record that has passed through the data linkage process and was not linked to any other record.



# Exegesis

---

## Thesis Abstract

### Background

Health and care systems are complex with many interactions and linkages. Using a whole system approach acknowledges that various interrelated factors can impact different parts of the health system and that service design and solutions to problems have to be developed taking all variables and interactions into consideration.

No single part of the health system provides the complete picture. It is the whole system and the way that the different parts work together that unlocks the research power of administrative datasets. The information to support whole system research requires the development of new approaches to join datasets and identify individuals across different data collections.

### Record Linkage Methods

In situations where recorded identifying information contains no error, and personal circumstances do not change, all that is necessary to link records from different parts of the healthcare system is to organise the individual records to be matched by personal identifiers. However, perfect datasets are rare, and it is more common that there will be discrepancies in identifying information between pairs of records belonging to the same person. In these situations, exact matching using these personal identifiers miss a significant number of correct ('true') links.

Alternative linkage methodologies have been developed to enhance data matching quality (i.e. to maximise 'true' matches and minimise 'false' and missed links). Traditionally, linkage systems use a combination of approaches to optimise linkage efficiency and accuracy. Techniques involved include deterministic approaches which use a combination of algorithms and rules to determine when two or more records match and probabilistic matching methods which use statistical theory to quantify levels of agreement and disagreement to make a decision whether two records belong to the same individual.

## Study Aims

The research aims to identify technical and functional requirements, as well as the motivations, for developing robust national linkage infrastructure in Australia. This thesis explores:

- The existing information that is known about large-scale linkage and knowledge gaps;
- The technical challenges associated with undertaking national record linkage; and
- How this research will extend the body of existing work and support researchers.

This thesis investigates record linkage as a method of creating population-level health information. The research defines how data from a diverse range of health datasets can be accurately and efficiently linked across jurisdictions and sectors, to enable nationally and internationally significant population-level studies.

## Results

Methods of matching have been developed and refined in Canada, England (Oxford), Scotland and Australia over the last thirty years that integrate data from different parts of the healthcare system and allow for imperfections in data.

In Australia, linked information has been embedded into the work of many state Governments and is used as a tool in decision making across health services and beyond. The traditional routinely linked dataset in Australia contains hospital discharges, cancer registrations and Registrar General's death records. Western Australia was one of the first regions in the world to build and routinely operate linkage infrastructure which provides a linked resource for state government and university researchers. Developments like this, and similar international examples, are used as a case study to understand the challenges associated with large-scale linkage and how the resulting permanent linkage infrastructure has been used to improve health, wellbeing and enhance the effectiveness and efficiency of health services.

The Population Health Research Network (PHRN) in Australia is a project funded through an allocation from the National Collaborative Research Infrastructure Strategy (NCRIS). The PHRN is a unique initiative which builds on the achievement of the Western Australian Data Linkage System (WADLS) and the NSW Centre for Health Record Linkage (CHeReL). The project provides a national infrastructure for the development and promotion of data linkage for population and clinical health datasets for research purposes.

The Centre for Data Linkage (CDL) is an essential element of the PHRN initiative. The CDL has created tools that enable linkage of data from a diverse and rich range of health datasets across jurisdictions and sectors (using demographic data) to facilitate nationally and internationally significant population-level research, using lessons learned from the record linkage community nationally and internationally. The resulting linked research datasets have no equal worldwide in terms of size and coverage.

## Conclusion

The benefits of data sharing have been shown to improve research skills and analytical tools for complex “linked data”, enabling new research that enhances the delivery of public services. University-based researchers in Australia are recognised as world leaders in the use of linked data for research.

With the development of the national linkage infrastructure, Australia now has a dedicated capability for the linkage of administrative and research data between States and Territories. Through the PHRN initiative, the CDL has created enterprise level linkage infrastructure that provides a platform for undertaking large, national linkage projects while meeting the requirement to provide a secure and controlled environment for working with sensitive data. The infrastructure has been designed to scale as dataset size and demand for national linked data increases. The infrastructure has considerably enhanced Australia’s ability to conduct high quality, internationally competitive research.

In a world where the growth in digital information and systems continues to expand, governments and researchers have access to unprecedented amounts of data. These large and complex data reservoirs require creative, innovative and scalable tools to unlock the potential of this ‘big data’. Record linkage is a powerful tool in the ‘big data’ arsenal. Linking (and integrating) data enables researchers to understand the complexities of systems, diseases and behaviours over time and to understand the needs of individuals, families and communities better.



## Aims and Objectives

Although state based linkage facilities have been supporting research in Australia for over thirty years, the primary goal of this research was to identify, understand and resolve the technical and methodological challenges associated with increasing the capacity of current record linkage facilities to create national linked data resources. The associated research focuses on maximising the value of data collected (both administrative and research) across institutional, geographical and portfolio boundaries using secure and robust methods.

The final goal was to explore how best to apply the techniques to large multi-state health service research projects. For this, a large proof of concept project was undertaken to investigate survival after hospital admission across four Australian states.

The following outlines the specific aims and objectives of this thesis.

**Aim 1            Understand the challenges and advantages associated with building and operating national linkage infrastructure.**

Objective 1    Review and compare developments, methods and models in operational national linkage systems so as to understand the technical and methodological challenges associated with creating national linked data resources.

Objective 2    Demonstrate the benefits of a dedicated national linkage resource to the research community. Explore the impact of routinely available linked resources on government and university research.

**Aim 2            Establish a national linkage operational model and governance framework to ensure privacy, security and confidentiality around linkage services.**

Objective 3    Determine an appropriate linkage and governance model for an operational national linkage facility.

Objective 4    Obtain approvals for the linkage and governance model from relevant jurisdictional linkage stakeholders and Human Research Ethics Committees (HRECs).

**Aim 3            Investigate, design and build efficient national data matching infrastructure for Australia.**

- Objective 5 Define a framework which can evaluate and benchmark linkage 'system' methodologies using appropriate performance metrics.
- Objective 6 Evaluate available matching system developments to identify strengths and avoid limitations.
- Objective 7 Describe the functions and features required in an effective national linkage system.
- Aim 4 Address privacy and security around record linkage: Privacy and security are important factors as links are traditionally established using demographic data held both within and across jurisdictions.**
- Objective 8 Identify appropriate technical and governance methods which can be used to minimise privacy concerns associated with record linkage.
- Aim 5 Investigate whether intelligent information technology can be used to measure and improve linkage quality.**
- Objective 9 Develop linkage quality metrics that will help to measure and understand the complexities of linked data.
- Aim 6 Validate national linkage facility and its ability to support national research.**
- Objective 10 Demonstrate the value of linkage methodologies and infrastructure in support of national research.
- Objective 11 Translation of results: show how record linkage and linked data can be used to predict issues, to develop practical solutions as part of planning processes inform (clinical) guidelines and change (clinical) practice.



## Thesis Overview

This thesis consists of peer-reviewed scientific publications, presented as a cohesive body of research, to demonstrate the utility of data matching methods as an effective piece of national infrastructure for population-level research.

### Chapter 1 - Record Linkage: an overview of the methods and developments around data matching and data sharing around the world

Chapter One provides an introduction to record linkage. This chapter describes the underlying methodologies used in record linkage and the elements involved in applying these techniques. These methods are developed, scaled and extended to national linkage projects as part of the research.

This chapter also addresses the first objective and provides an international case study which looks at challenges associated with developing the Scottish record linkage infrastructure. This routinely operated national linkage system has been developed and improved since the 1980's. The vision for the Scottish system has always been to create and maintain national linked datasets using automated algorithms with no intervention involved. This principal is important for large national systems which have finite resources to manage and quality check the 'big data' involved in these linkages. Many of the lessons learned have influenced the investigations, methods and designs used to develop national infrastructure in Australia. (1 supporting manuscript).

Research in this chapter is covered by the following peer-reviewed scientific publication(s):

14. Walsh D, Smalls M, **Boyd J**. *Electronic health summaries--building on the foundation of Scottish Record Linkage system.*(2001) *Medinfo, 10 (Pt 2)*, pp. 1212-1216.

### Chapter 2 - A review of the technical and methodological challenges associated with creating national linked data in Australia.

Chapter Two addresses the second aim, to establish national linkage infrastructure in Australia. One of the main tasks was to develop an operational model and governance framework which was acceptable to stakeholders. A national linkage methodology has been developed using methods which separate demographic variables such as name and address from clinical variables including health and services variables. This approach is often

referred to as “the Best Practice Protocol” (Kelman, Bass and Holman 2002). As well as separating the functions associated with linkage, the infrastructure model also addresses the appropriate information and security standards using a robust governance framework which was developed by, and agreed upon with the input of key stakeholders (including linkage units in Australian States and data custodians).

This chapter also addresses the third aim to evaluate and build efficient national record matching infrastructure for Australia. To efficiently and effectively evaluate linkage methods and software, a transparent and transportable evaluation methodology was developed. Using synthetic datasets and standard performance metrics, this allows users to approach linkage performance issues without breaching privacy concerns. The synthetic data and metrics have been recognised as significant research outcomes and are being used nationally and internationally to benchmark linkage methods.

The operational model and lessons learned from other ‘big data’ systems were used to inform the development of a national linkage system. The national linkage infrastructure creates person-based linkage keys across multiple nodes using common demographic variables. These inter-node linkages can be used to explore cross-border flows, which anecdotally have been believed to be significant, but have never previously been quantified. System design has been carefully researched to ensure it includes all features required by an ‘enterprise’ facility. (3 published manuscripts).

Research in this chapter is covered by the following peer-reviewed scientific publication(s):

1. **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 services research 12.1 (2012): 480.
2. **Boyd JH**, Randall SM, Ferrante, AM Bauer JK, Brown AP, Semmens JB. Technical challenges of providing record linkage services for research (2014) BMC Medical Informatics and Decision Making, 14 (1), art. no. 23.
7. Ferrante AM and **Boyd JH** (2012) A transparent and transportable methodology for evaluating Data Linkage software. Journal of Biomedical Informatics (45)165-172.

### Chapter 3 - Privacy and data linkage

Chapter Three addresses the fourth aim by exploring privacy and data confidentiality issues. Record linkage raises issues of privacy and confidentiality as it requires personal identifying

information to accomplish matching, and the resulting linked data provides a broader (more comprehensive) picture of the individuals involved. The research project identifies and assesses Privacy Preserving Record Linkage (PPRL) techniques on large real-world datasets. (1 Book Chapter and 1 published manuscript).

Research in this chapter is covered by the following peer-reviewed scientific publication(s):

12. **Boyd JH**, Randall SM, Ferrante AM. Application of privacy preserving techniques in operational record linkage centres. *Medical Data Privacy Handbook*. Springer International Publishing, 2015. 267-287.
9. Randall SM, Ferrante AM, **Boyd JH**, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. *Journal of Biomedical Informatics* 2014; DOI: 10.1016/j.jbi.2013.12.003.

#### Chapter 4 - Methods to assess linkage quality - how do we measure up?

Automated and semi-automated methods to improve linkage quality are investigated in Chapter Four to address the fifth aim. Traditionally, clerical monitoring shows on a pair-wise basis, both the false positive rate (the proportion of records which are incorrectly linked) and the false negative rate (the proportion of records which the system fails to link). By developing quality metrics, automated break links and using a focused approach to clerical checking, the project develops techniques which achieve the quality advantages of a fully clerically checked system without the massive investment of time and expense such systems typically require. (3 published manuscript and 1 published letter)

Research in this chapter is covered by the following peer-reviewed scientific publication(s):

3. **Boyd JH**, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, Dickinson T. A simple sampling method for estimating the accuracy of large scale record linkage projects. *Methods of information in medicine*. 2016;55(3):276-83.
10. Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making* 2013; 13 (64): e1-e10.
11. Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. Use of graph theory measures to identify errors in record linkage. *Computer Methods and Programs in Biomedicine*. Volume 115, Issue 2, July 2014, Pages 55-63.
13. **Boyd JH**, Ferrante AM, Irvine K, Smith M, Moore E, Brown AP, Randall SM. Assessing linkage quality - what do researchers need to know? *Australia and New Zealand Journal of Public Health* doi: 10.1111/1753-6405.12597.

## Chapter 5 - National Data Linkage - Proof of Concept

Chapters Five deals with the final aim. Having developed an accurate, reliable, load-bearing (i.e. production capability) record linkage infrastructure, the environment was tested through the first Proof of Concept project (PoC#1). PoC#1 involved person-level linkages of hospital admission and death records across four jurisdictions (NSW, WA, QLD and SA) for the ten year period 2000 through 2009 (over 45 million records). The linked dataset created for this project is one of the largest ever constructed worldwide, with more records and matches than most established routine national and international linkage systems. (2 published manuscripts).

Research in this chapter is covered by the following peer-reviewed scientific publication(s):

4. **Boyd JH**, Randall SM, Ferrante AM, Bauer JK, McInnery K, Brown AP, Spilsbury K, Gillies M and Semmens JB. Accuracy and completeness of patient pathways - the benefits of national data linkage in Australia. *BMC health services research* 15.1 (2015): 312.
6. Spilsbury K, Rosman D, Alan J, **Boyd JH**, Ferrante A, Semmens JB. Cross border hospital use: An analysis using data linkage across four Australian states. *Medical Journal of Australia* 202.11 (2015): 582-586.

## Chapter 6 - Using National Record Linkage Infrastructure to Support Research

Chapter Six The final chapter looks at the impact of accessible linkage infrastructure on the research community. Significant and influential research programmes in both Scotland and Western Australia are presented which demonstrate the benefit of routinely available linked resources for the research community (Objective 2).

Further, this chapter explores the impact of Scottish record linkage on the research community. This unique and powerful resource has been widely used in Scottish research projects for over thirty years, with demand increasing year on year. Some of the early work with the Scottish linked data has seen the development of national clinical outcomes and disease profiles which are now part of routine reporting in national statistics. (4 supporting manuscripts)

In addition, the chapter describes the outputs from the Western Australian Population-based Burn Injury Project (WAPBIP): a retrospective cohort investigation using Western Australian population-based linked data for burn and uninjured groups. The overarching aim of the

project is to provide important information to inform burn care, prevention, education and policy both nationally and internationally. The study uses linked health data to explore patient pathways, hospital utilisation and costs of burn injury over the last thirty years. (2 published manuscripts).

Both these studies have been used to help plan clinical services, inform clinical practitioners and change clinical practice.

Research in this chapter is covered by the following peer-reviewed scientific publication(s):

5. **Boyd JH**, Wood FM, Randall SM, Fear MW, Rea S, Duke, JM. Effects of pediatric burns on gastrointestinal diseases: A population-based study. *The Journal of Burn Care & Research* (2016). (In Press).
8. Duke J M, Bauer J, Fear MW, Rea S, Wood FM, **Boyd J** (2014). Burn injury, gender and cancer risk: population-based cohort study using data from Scotland and Western Australia. *BMJ open*, 4(1), e003845.4
15. Capewell S, Kendrick S, **Boyd J**, Cohen G, Juszczak E, Clarke J. Measuring outcomes: One month survival after acute myocardial infarction in Scotland. (1996) *Heart*, 76 (1), pp. 70-75.
16. Capewell S, MacIntyre K, Stewart S, Chalmers JWT, **Boyd J**, Finlayson A, Redpath A, Pell JP, McMurray JJV. Age, sex, and social trends in out-of-hospital cardiac deaths in Scotland 1986-95: A retrospective cohort study (2001) *Lancet*, 358 (9289), pp. 1213-1217.
17. MacIntyre K, Capewell S, Stewart S, Chalmers JWT, **Boyd J**, Finlayson A, Redpath A, Pell JP, McMurray JJV. Evidence of improving prognosis in heart failure: Trends in case fatality in 66 547 patients hospitalized between 1986 and 1995 (2000) *Circulation*, 102 (10), pp. 1126-1131.
18. MacIntyre K, Stewart S, Capewell S, Chalmers JWT, Pell JP, **Boyd J**, Finlayson A, Redpath A, Gilmour H, McMurray JJV. Gender and survival: A population-based study of 201,114 men and women following a first acute myocardial infarction (2001) *Journal of the American College of Cardiology*, 38 (3), pp. 729-735.



## Chapter 1

---

### Record Linkage: an overview of the methods and developments around data matching and data sharing around the world

*“If you can't explain it simply, you don't understand it well enough”*

*Albert Einstein*

#### **Supporting Manuscript(s):**

Walsh, D, Smalls M, and **Boyd J**. *Electronic health summaries-building on the foundation of Scottish Record Linkage system*. *Studies in health technology and informatics 2* (2001): 1212-1216.





## 1.1. Introduction

The magic surrounding record linkage was first portrayed by *Dunn* in 1946 as a 'Book of Life': for each person, life starts with a record of birth and ends with a record of death [1]. This description of a volume containing a chronological history of significant life events from every aspect of a person's lifetime provides a perfect picture of what record linkage can achieve, with each book containing a different story.

Even before the power of computers was harnessed to progress record linkage as a discipline, *Dunn* emphasised the importance of accurate and complete information which would work together to enhance an individual's story. The importance to research was evident, cross referencing vital events from a population would provide meaningful patterns which could be used by public health and statistical agencies to improve community health and welfare [2-4].

Record linkage was initially undertaken using manual references and punch cards which allowed basic analysis of linked data. In the 1980s, computer power provided the significant move from small studies to large population based linkage [5]. Since the 1980s the information revolution has seen digital information about our lives grow exponentially. This 'Digital Era' is characterised by technology which increases the volume, variety, velocity and the veracity of data about individuals and the society in which they live [6].

Government and University departments around the world appreciate that linked administration data can provide a unique resource for monitoring, evaluating and improving services [7-9]. However, due to various technical and legal barriers, it has not always been possible to make these data available to researchers. In Australia, various administrative datasets are gathered at different tiers of government (Federal, State and Local Government). The data collected by these organisations is not readily available to one authority making 'joined up' research of government services very difficult.[10]

## 1.2. Record Linkage Methodology

In situations where identifying information is recorded without error, and personal circumstances do not change, the matching process can be reduced to a simple sort of the records by personal identifiers. However, perfect datasets are rare, and it is more common that there will be discrepancies in identifying information between pairs of records belonging to the same person. In these situations, exact matching using these personal identifiers miss a significant number of true links.

Alternative linkage methodologies have been developed to enhance data matching quality (i.e. to maximise 'true' matches and minimise 'false' and missed matches. Traditionally, linkage approaches use a combination of approaches which involve deterministic and probabilistic matching methods [11].

Deterministic matching systems apply a series of business rules to decide whether two (or more) records belong to the same person (these rules based algorithms "determine" the result). In these systems, the linkage results are strictly decided by the business rules i.e. either a linkage comparison meets the defined business rule or it does not [12].

Files with high quality data and many variables are frequently linked using a deterministic approach (i.e. generate links based on exact agreements among individual identifiers). When the number of data attributes and rules required is small, the development of the deterministic matching algorithms is relatively simple and is easy to implement. Where the linkage involves large datasets with complex characteristics, the more complicated the rules-based matching routines become [13].

In most administrative data collection systems, the datasets are large increasing the potential for duplicates, human error and discrepancies. The system design must allow for complex error patterns within true links enabling us to determine links within and between data files. Deterministic matching systems are typically less sensitive to errors/discrepancies in the data and as a result will miss more links compared to a probabilistic approach to matching. A deterministic linkage method is most applicable when the number of records to be matched is relatively small, there are a limited number of data attributes for linkage, and there are minimal recording errors within the underlying datasets [14].

Probabilistic systems are based on a statistical model and use likelihood ratio theory to assess record pairs, quantifying the probability that two records belong to the same entity. This method does not rely on exact matching across the data attributes and can more accurately establish links between records with more complicated error patterns made in the recording process than deterministic systems [15-17].

Frequency distributions of the data available for matching can be used to determine an optimal probabilistic strategy for linkage [18, 19]. These systems can be easily amended to accommodate a growing number of data files without significantly impacting on performance or linkage accuracy. The statistical approach to linkage makes it relatively easy to optimise, implement and maintain a probabilistic matching strategy over time [17, 20, 21].

For situations where dataset sizes and numbers of attributes are large, high levels of accuracy and low total cost are important; organisations should select a probabilistic system. When datasets are smaller, have fewer attributes and accuracy is not a major factor, then a deterministic approach may be preferable [14].

### **1.3. Components of the data linkage process**

Notwithstanding the size of the datasets and the methods of linking, matching the records consists of carrying out the same basic operation. This process involves the comparison of two records and the decision as to whether they belong to the same individual [16, 22, 23]. The linkage process is implemented in a number of steps:

#### **a) Blocking - Selecting pairs of records together for comparison**

In an ideal world, to maximise linkage quality, we would carry out matching between every possible pair of records to determine whether they belong to the same person. For large files, it is not practical to conduct matching on all pairs of records involved in a linkage.

To cut down the number of pair comparisons, only a subset is compared. We compare only those records which share a minimum level of identifying information. This reduction was traditionally achieved by sorting the files into 'blocks' within which paired comparisons are carried out [11, 24].

In deterministic matching, the matching rules can be equated to blocking strategies i.e. any record-pair identified by the set of rules are considered a matched record-pair. In contrast, within a probabilistic approach, records will be compared if they meet the criteria specified across each of the blocking rules but depending on the level of agreement they may or may not be designated as a match [12, 25].

It is, of course, possible that two records belonging to the same person will not meet the blocking criteria and never be eligible for comparison. It is important that the set of blocks are defined in a way that minimises links lost because of blocking without comparing too many 'true negatives' which is computationally expensive [26].

#### **b) Matching**

Linkage systems often use a mixture of approaches which involve using a combination of rules (deterministic) and statistical theory (probabilistic) to determine matches [27].

Deterministic methods are computationally inexpensive (relative to probabilistic methods) and are easier to implement [12].

Probability matching allows a mathematically precise assessment of the levels of agreement and disagreement between two records [28]. Having identified candidate pairs, two principles are applied during the probability matching:

- Every time an element of identifying information on two records is the same, the probability that they belong to the same person is increased.
- Every time an item of identifying information differs between two records, the probability that they apply to the same person is usually decreased.

When we compare items of identifying information between two records, we obtain an 'outcome'. This outcome will increase or decrease the level of agreement between the two records. The level of agreement or disagreement is based on the following two questions:

- How often is this outcome likely to occur if the two records belong to the same person (the comparison is a true positive)?
- How often is this outcome likely to happen by chance i.e. if the two records do not belong to the same person (the comparison is a true negative)?

The ratio between these two probabilities is called an odds ratio. The odds ratio quantifies how much a particular 'outcome' increases or decreases the likelihood that two records being compared belong to the same person [29]. Where possible, specific weights relating to degrees of agreement and disagreement are calculated based on the data being matched. In theory, it is feasible to include any items of identifying information between two records if they have an influence on the chance that the two records belong to the same individual [30, 31]. However, it is important that items included in the matching algorithm are, as far as possible, statistically independent [16, 32].

### **c) Finding matches**

Whatever kind of matching is being performed, whether linking records within a file or linking records between files, the decision making process is the same, examining pairs of records and making a judgment about whether they belong to the same person. Overall, the matching aims to divide all the pairs into two classes which are more generally referred to as 'true positives' or 'true negatives'.

The linkage methodology tends to be customised for each dataset with the aim of maximising the ability to detect 'true positives' i.e. correct matches using the available identifying information. In deterministic matching, rules are designed around the available identifying variables within the datasets. Probabilistic methods calculate particular scores for each pair comparison. As a result, the distribution of probability scores differs for each kind of linkage. The crucial step in each linkage is to identify an accepted threshold (cut-off) above which a pair is considered a match (and below which it is considered not to be a match) [33].

The threshold (cut-off) is typically decided by manual inspection of a sample of the pairs of records. Comparison scores above the threshold indicate that it is more likely than not that the two records refer to the same entity. A comparison score below the threshold suggests that the two records do not belong to the same person [32].

Once a threshold has been set, the systems will apply the matching strategy and decide whether records belong together. In practice, development of the matching algorithms and setting the threshold is an iterative process with results affected by a range of factors; including the quality of the data and the characteristics of the datasets involved.

#### **d) Grouping records**

The final step is to create groups of record belonging to the same person from the record pairs. Grouping strategies amalgamate collections of record pairs found through the matching process, to determine the full set of records belonging to the same individual. This process makes use of the ordinal properties of the matched record-pairs. In probabilistic linkage, the odds weight (score) of each record pair is used, with higher weighted pairs deemed more likely to be a correct match. In deterministic (rules based) linkage, the order in which the rules are applied is used as a marker of quality – more stringent rules are applied first, and record-pairs created through these rules are deemed more likely to be a correct match than those arising from later, looser rules [34].

The type of grouping strategy used is closely related to the properties of the data being linked. Of particular importance is whether each dataset is expected to contain multiple records for a single person, or only one record per person [35].

With increased processing power and data storage capacity, it is possible to operate a system in which all records for an individual can be linked once and held together on a permanent dataset. Once data linkages have been made and then preserved, it means that

the cost of linkage does not have to be incurred again for each new project that requires linked data and so will be more cost effective in the long run [36, 37].

## **1.4. National and International record linkage developments**

Despite increasing use of linked data by university and government researchers, dedicated record linkage infrastructure routinely running linkages to support data linkage activity is still limited. Record linkage “systems” or “facilities” exist in only a handful of countries including Canada [38], England (Oxford) [22], Scotland [39], Australia [40, 41] and most recently in Wales through the development of the SAIL system [42]. These enterprise-level facilities operate routinely, undertaking linkage to support statistical and research needs of the government and research community.

### **1.4.1. The Oxford Record Linkage Study (ORLS)**

The Oxford Record Linkage Study (ORLS) was established in 1963 by Donald Acheson [43, 44]. The study, funded primarily by the NHS, started as a joint project involving the National Health Service (Oxford Regional Health Authority (RHA)) and researchers (University of Oxford). The rationale behind the ORLS was to maximise the value of existing data by making linkage possible for epidemiological and health services research in particular by using NHS statistical data and cohort methodologies.

Following the decommissioning of RHAs in 1995, the ORLS relocated to the University unit and continued to gather hospital data from health authorities within the former Oxford RHA. From 2005, the NHS National Information Centre created linked English national data for a variety of research topics with funding from the Department of Health. More recently, with funding from the National Institute for Health Research, the group in Oxford continues to take the Oxford subset for ORLS and supports linkage of the national English data [45, 46].

### **1.4.2. The Manitoba Population Health Information Systems**

The Manitoba Centre for Health Policy (MCHP) is located within the Faculty of Medicine at the University of Manitoba. Since 1974, Manitoba Health has provided health care utilisation data to University of Manitoba researchers. In 1991, with the establishment of the MCHP, this information was placed in the MCHP repository. Since 1991 the number and datasets in the repository have grown quickly allowing research on the health of Manitobans.

The linkage maps (based on the hashed Manitoba Personal Health Identification Number) are created in the Health Information Management Unit of Manitoba Health using identifying information provided by ‘Data Trustees’ (using a combination of health numbers or

deterministic/probabilistic linkages on personal identifiers where health numbers are not present). The anonymised linkage maps are provided to MCHP to allow integration for approved research projects. The anonymised administrative data derived from administrative claims are stored as separate unlinked files by MCHP and integrated using the linkage map (a hashed identifier called the Manitoba Personal Health Identification Number and unique record numbers).

The Manitoba Health Registry sits at the repository's core, with a universal healthcare system, almost all residents have a Manitoba Health card. This provides almost complete population coverage and an excellent basis for both linkage and related research [47].

### **1.4.3. Population Data British Columbia**

Population Data BC (PopData) is a collaborative university research infrastructure capability which was established in 2009 to maintain and enhance the British Columbia Linked Health Database (BCLHD). BCLHD was set up by the Centre for Health Services and Policy Research at the University of British Columbia in 1996. PopData facilitates research from data holdings which have been extended from traditional health data (BCLHD) to include data files from education, childhood development and the environment.

Operating as a trusted third party for record linkage, PopData supports approved research access to person based, de-identified longitudinal linked data collections from British Columbia's 4.4 million residents. The research based linked datasets for research include physician payments, PharmaCare, hospital separations, continuing care, birth registrations, death registrations, mental health episodes of care, early childhood data, Worker's Compensation Board and the British Columbia Cancer Agency cancer incidence and spatial data [48].

### **1.4.4. Secure Anonymised Information Linkage (SAIL)**

SAIL (Secure Anonymised Information Linkage) databank is curated by the Health Information Research Unit (HIRU) of the School of Medicine at Swansea University. HIRU aims to maximise the value of routinely collected individual level data through record linkage and to enable and support health related research by government and the wider research community [42]. HIRU works in partnership with researchers and health professionals to support clinical research, patient outcomes, service delivery and health improvement.

Linked data from different sources is created by the National Health Service (NHS) in Wales using the NHS number. The NHS Administrative Register (NHSAR) provides personal

information of all persons who have registered with a GP practice or received care from health services in Wales. The SAIL system employs a data separation policy to ensure privacy of individuals; this involves separating data into clinical and demographic components by data custodians at the source organisation and allocation of an anonymous field/key to enable linkage between data files. [49]

#### **1.4.5. The Western Australia Data Linkage System (WADLS)**

Linking data collections for medical research and health service planning has a long history in Western Australia. The origins of the WA system can be traced back to the development and achievements of the ORLS through Emeritus Professor Michael Hobbs [50]. Following very successful record linkage projects during the 1970s and 1980s, formalisation of record linkage infrastructure in Western Australia was established through a collaboration between the Department of Health Western Australia (DoHWA), Curtin University, the University of Western Australia (UWA) and the Telethon Kids Institute (TKI) [51].

The Western Australia Data Linkage System (WADLS) was established in 1995 initially through Western Australian lottery funding to connect all available health and related information for the WA population. WADLS is a production probabilistic data matching system which creates, stores, updates and extracts links between over 40 population-based administrative and research data collections in Western Australia [51].

The WADLS is operated by a team located within the Western Australia Department of Health which routinely links core data about health events across all individuals in Western Australia. The links are created using the 'separation model' developed in Western Australia to provide additional privacy protection to operations [52]. Under this model, the data custodians separate personal identifiers from clinical data for linkage.

To provide a complete service for researchers wishing to access the data, the data linkage branch also provides:

- Client services - to manage linkage projects and support research applications for linked data;
- Data linkage - to perform the linkage; and
- Analysis and quality – to operate the Custodian Administered Research Extract Server (CARES).



These functions combine to provide support through the application and approvals process as well as the technical components of linkage. The resulting linked information is used in research, planning and evaluation projects which have appropriate ethical approval and whose aim is to improve the health of Western Australians.

#### **1.4.6. The Centre for Health Record Linkage (CHeReL)**

The Centre for Health Record Linkage (CHeReL) was established in 2006 to create and sustain record linkage infrastructure for the health and human service sectors in NSW and the ACT that would provide a mechanism to access linked data for researchers, health planners and policy makers.

The CHeReL's role includes both linking these datasets together to determine which records within and between datasets belongs to the same person, as well as servicing clients with this information. The CHeReL data runs from 1994 and comprises over 84 million records (with approximately 10.4 million people). This linkage process involves both deduplication of datasets and linkage between datasets [41].

The CHeReL core system (the Master Linkage Key) consists of NSW hospital, emergency, perinatal and mental health datasets, along with the NSW cancer registry, birth and death registries, and the notifiable conditions registry. Similar datasets for ACT also form part of the CHeReL core system. In addition to these core datasets, CHeReL regularly links additional datasets as required for particular research projects.

CHeReL has developed large scale linkage infrastructure to create and manage the Master Linkage Key (MLK). The CHeReL system consists of several components, these include:

- A data management system (DMS) that manages data both before, during and after linkage;
- A linkage engine (ChoiceMaker) that determines which records belong to which individual; and
- Quality assurance procedures, including manual review processes.

The CHeReL system manages routine 'ongoing' incremental linkage - that is, incoming records are linked both to themselves and to all existing records within the system which have already been given person identifiers.

Along with the linkage system and standard operating procedures, CHeReL has a programme in place to look at the performance of the linkage system (especially as the Master Linkage Key and the demand for linked data continue to grow). This focus on continuous improvement is an ongoing effort to improve the quality of linkage services, processes and outputs. Enhancements to the managed linkage system allow CHeReL to provide accurate, up-to-date linked information which is responsive to customer needs as the demand for linkage increases.

The CHeReL production system has been custom built to manage the whole linkage process (from beginning to end) and to maintain the resulting Master Linkage Key (MLK). The system uses an integrated linkage engine (ChoiceMaker) [53].

The system model includes the separation of functions within processes that create the CHeReL MLK. Processes to match records and build the MLK are highly integrated. MLK datasets (tables) are structurally relationally and stored in a SQL Server database. The documentation around system operation and MLK processes appear comprehensive.

The CHeReL linkage system is one of the largest production linkage systems in the world supporting a population base of over 7 million people. As a result, the CHeReL operating model is necessarily large and complex to accommodate the numerous multifaceted health and health-related data collections used to support approved research projects.

## **1.5. The first routine national record linkage system**

### **1.5.1. Record Linkage in Scotland**

Although record linkage has established itself as a crucial element within the field of 'Big Data' science, there are very few facilities internationally that can provide routine data matching at a national level. Often record linkage infrastructure is based and operated around individual geographical, portfolio and legislative jurisdictions rather than servicing national requests. This can lead to disparate operations providing linkage services within silos without the option for interoperability.

Scotland is one of the few countries that has established large scale routine national record linkage infrastructure to support both university and government research [54]. In order to better understand the challenges associated with building and operating national linkage infrastructure, the project explored the development of record linkage in the National Health Service (NHS) in Scotland as a case study [55].

NHS National Services in Scotland collects, manages and stores a wide variety of Scottish health data required to monitor and report on health services routinely. Like many other countries, the statistical services of the NHS and Scottish Government use these valuable information resources for national reporting, to plan services and to ensure efficient and effective delivery of patient care.

As part of the evolving national administrative data collections, important decisions were made in the late 1960's enabling Scotland to embark on national medical record linkage. The decision to collect names on hospital administrative returns was made with record linkage and patient based analysis in mind. As a result, all hospital discharge records, cancer registrations and death records from 1968 are held centrally in machine readable form and contain patient identifying information (name, date of birth, gender and area of residence) [56, 57].

The idea that administrative medical records could be brought together on a patient basis across the whole country was first outlined by Heasman in 1968 [57, 58]. This iconoclast laid the foundations for the Scottish Medical Record Linkage system based on his knowledge of the early linkage work carried out in Canada and Oxford [44, 59, 60]. The establishment of a small Record Linkage team in 1968 allowed Scotland to build a track record and to remain at the forefront in linking data for research purposes for half a century [39].

### **1.5.2. Early development in Scotland**

Initially, new linkages were carried out for each project, and each linkage involved bespoke matching algorithms (each of these linkage projects took between 6 months and a year to complete). As the power of computers and data storage capacity increased it was possible to consider developing a routine linkage system with enduring links.

The 'production' Scottish Medical Record Linkage System has been operating for over 25 years and brings together all health records belonging to the same person as patient groups stored within permanently linked datasets. NHS National Services in Scotland maintains two permanent linked files, the Scottish Morbidity Database with records dating back to 1980/1981 and the Maternity and Neonatal Database containing obstetric discharge records for all mothers delivering in Scotland since 1975, as well as related baby records. The Scottish Morbidity and Maternity and Neonatal Databases are updated monthly and six monthly respectively with new incoming records linked into the existing databases using probability matching [39, 61, 62].

With a permanently linked file, it is easier to extract and provide research data avoiding duplication of effort by data custodians and linkers (who were previously linking the same data over and over for different projects). With this linked resource a permanent feature, it also became possible to routinely research and report health service activity using patient based analysis in National Statistics.

Initially, the system was updated annually with the addition of a new year of each of the data types. This annual linkage used the traditional sort and matching techniques; this process took around two months elapsed time on the mainframe [35, 63]. The linkage took place around June in the year following the year for which the data was to be added, allowing time for final validation of the calendar year of unlinked data.

The linkage schedule involved between six and eighteen months delays before linked data was updated for a new calendar time period and made available for analysis. The problem with this traditional method of linking is that it involves sorting all the records in a variety of ways to optimise the number of links found [35]. When linking a relatively small number of new records to a master linkage file containing millions of records, the sorting procedure becomes more resource intensive, to the point of being unsustainable, as the master linkage file grows.

### **1.5.3. Year 2000 (Y2K) redevelopment**

As part of the Y2K system redevelopment, NHS National Services in Scotland dedicated resources to design and build an efficient system which would link national data files frequently. The project aimed to provide accurate, up-to-date linked information which was more responsive to customer needs.

The new Production Linkage System was designed with the enhanced functionality of relational database technology; enabling 'one-step' linkage, by continuously updating the linked database during linkage and thus avoiding the need for the extra stage of 'internal' linking of 'newcomer' records.

The linkage process was also split into two phases:

- Exact Match
- Probability Match

The exact match process accounts for up to 40% of 'new' records which can be treated as secure links, without a calculation overhead the exact match runs faster than the traditional

probability match. The records that remain unmatched from the exact match phase pass into a probability match. This stage uses a probability matching methodology and has the advantage of matching a smaller file which speeds up the process.

#### **1.5.4. 'One-pass' Linkage and the Best Link Principle**

To avoid limitations of the original system which sorted the whole linked file as part of the processing, the Y2K system stored new 'incomer' records in memory. The existing linked records could then be read in sequentially and compared with all new records which fit the chosen blocking criteria.

In order to take advantage of existing linkage information, records in the linked files are read in as patient record sets where appropriate. Therefore, the decision whether to link two records depends not only upon the probability weight achieved by the comparison of two records but also on the other probability weights achieved by any of the records in the patient record set. This is known as the 'best link' principle [64].

In practical terms, best link means that the system does not have to interfere with the linkage structure of the data already in the linked dataset. The system simply assigns to incoming records the group number of the linked records with which they achieve the best link. The process does not allow any existing groups of records to be joined because they have both linked to the same new record (i.e. it does not allow bridging).

#### **1.5.5. Community Health Index (CHI) number**

The Community Health Index (CHI) number is a unique ten digit number that the majority of people receive during the registration process with a GP practice in Scotland [65]. The CHI number is used as the primary patient identifier throughout Scotland and was crucial to eHealth as well as Information and Communication Technology (ICT) strategies in Scotland.

Based on the regionally maintained CHI, the Scottish Care Information (SCI) programme supported the development of Electronic Patient Records across the NHS [54]. The SCI programme also included the implementation and support for the standard use of a Unique Patient Identifier (UPI) as a single patient reference number across the whole of Scotland. The probabilistic linkage system within NHS National Services in Scotland was used to seed and maintain the operational Master Patient Index (MPI) with the UPI. The idea was to create a quality UPI which was utilised in all clinical communications removing the need for routine probability matching.

### **1.5.6. Linking health and social data in Scotland**

The current model in Scotland uses both the CHI and linked files to maximise matching efficiency in health systems and research projects. The production linkage algorithms have been amended to use the UPI as well as probabilistic matching. This enhanced matching model has increased efficiency and quality of the linked data.

NHS National Services in Scotland also provide the facility to seed the CHI number into a variety of health and other records using a probabilistic matching algorithm that identifies an individual's CHI number from their personal data such as name, gender, date of birth and address. It can determine CHI numbers with high accuracy, even where an individual's personal data is not necessarily up-to-date or entirely accurate. Increasingly, Local Authorities and other agencies are seeing the benefit of attaching the CHI number to records. In particular, this is useful for social work as the CHI number provides the means for case-specific information sharing with Health Boards, GPs and other agencies [66].

Where the CHI number is unavailable (e.g. historical data) traditional probability matching is used, and the record linkage unit within NHS National Services in Scotland is recognised internationally for its capability to link clinical and research databases to routine hospital admissions and death registrations.

### **1.5.7. Research and development around record linkage**

Primarily funded through the regular operating budget of NHS National Services in Scotland, there have been relatively little resources and effort available for research into enhancing the linkage methodology. Instead, the system has been developed and refined in response to the wide and varied requirements of operational demands.

The linkage system is a core component of the Scottish Health Service generating routine data for national statistics and supporting specialised projects requiring bespoke linkages within very tight deadlines. The Scottish record linkage system provides a benchmark for routine operations, and identifies current limitations within operational systems, this was used as the starting point for developing national linkage research in Australia and outlines challenges associated with the developments [67].

### **1.5.8. Conclusion**

This chapter explores and describes the challenges associated with developing an operational system that can deliver routinely linked data for Australia. It provides an

overview of the record linkage methodology and components used to develop national linkage infrastructure as part of the research. The chapter also addresses the first objective of the thesis, providing an international case study that describes the strengths and weaknesses of national linkage infrastructure in Scotland. The supporting paper provides an overview of the Scottish linkage system including the advancement from project to routine processing, elements of the linkage system, linkage outputs (with the building blocks) and benefits to the research community.

Lesson learned from the Scottish system have helped shape the system design, methodological investigation and operational methods used to develop national infrastructure in Australia.





## 1.6. Supporting Manuscript

Walsh, D, Smalls M, and **Boyd J.** *Electronic health summaries-building on the foundation of Scottish Record Linkage system.* Studies in health technology and informatics (2001)



## Chapter 2

---

### A review of the technical and methodological challenges associated with creating national linked data in Australia

*"Men are only as good as their technical development allows them to be"*

George Orwell

#### **Published Manuscript(s):**

**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 services research 12.1 (2012): 480.

**Boyd JH**, Randall SM, Ferrante AM, Bauer JK, Brown AP, Semmens JB. *Technical challenges of providing record linkage services for research* (2014) BMC Medical Informatics and Decision Making, 14 (1), art. no. 23.

Ferrante AM and **Boyd JH**. *A transparent and transportable methodology for evaluating Data Linkage software.* Journal of Biomedical Informatics (2012) (45)165-172.

#### **International Conference presentation(s):**

**Boyd JH**, Ferrante AM, Randall S, Bray J, O'Shea, A, Semmens JB. *Development of National Data Linkage: A linkage system for the 21st Century.* International Health Data Linkage Network Conference (IHDLN). Perth, Western Australia, May 2012

**Boyd JH**, Ferrante AM, Randall SM, Bauer, J, Gillies M, Semmens JB. *Developing National Data Linkage Infrastructure in Australia.* SHIP International Conference: Exploiting existing data for health research. St Andrews, Scotland, September 2013.



## **2.1. Data Linkage development in Australia**

The potential for a linked data system in Australia was first proposed by Hobbs in 1970 [50]. He outlined a vision for medical record linkage studies on a national or state-wide basis and suggested the introduction of a pilot linkage scheme in Western Australia (WA).

Data linkage systems were being developed in WA from the mid-1980s. The Telethon Institute for Child Health Research established a WA Maternal and Child Health Research Database, linking information from midwives notification records, birth registrations, death certificates, hospital inpatient morbidity data and the congenital disabilities and cerebral palsy registers to perform paediatric epidemiological studies. Today, this is incorporated into the WADLS [40, 51]. Another early project was the WA Road Injury Research Database, which links state-wide police, hospital and death records of road crash casualties. Another significant early use of data linkage in WA on an ad hoc basis was in the ongoing follow-up of men who mined asbestos at Wittenoom Gorge. However, lack of political will, computing and resource constraints precluded the development of full population-based data linkage in WA until 1995, when a large infrastructure grant to establish the existing system was awarded by the WA Lotteries Commission (Lotterywest) [51].

The idea of a permanent, dedicated linkage infrastructure for New South Wales (NSW) was realised in 2006 with the establishment of the Centre for Health Record Linkage (CHeReL). The linkage unit, managed by NSW Ministry of Health, has established relationships with a broad range of research and government organisations and has incorporated core administrative data collections for NSW and ACT into the Master Linkage Key (MLK). The MLK is well used by researchers, planners and policymakers in NSW and ACT [41].

Linkage of large population health datasets within these state linkage units has resulted in information which is routinely used to plan, implement and evaluate a range of health services and to identify areas for improvement. Linked population-level health data can be used for timely and cost effective evaluation of health care policy and service provision. Linked population health data has the advantage of being representative of the whole population, allowing efficient use of existing data resources, and are cost-effective compared to collecting data from large numbers of people [51, 68].

## **2.2. Population Health Research Network**

The Population Health Research Network (PHRN) was established to provide Australian researchers with access to linkable de-identified data from a diverse and rich range of health datasets, across jurisdictions and sectors. This will support nationally and internationally

significant population based research to improve health and enhance the delivery of health care services in Australia.

The PHRN received an initial \$20 million allocation from the Australian Government National Collaborative Research Infrastructure Strategy (NCRIS) program [69]. The funding period is 2008-09 to 2010-11. The Australian Government announced a further \$10 million for the PHRN in the May 2009 budget. The allocation was sourced from the Education Investment Fund. The additional funding enabled further enhancement of the infrastructure developed through the NCRIS program in 2011-12 and 2012-13. In addition, state and territory governments and academic partners contributed a further \$32 million in cash and in-kind contributions to the PHRN in the 2008-09 to 2010-11 period.

The PHRN has the objective of developing a national infrastructure and promotion of data linkage for population and clinical health datasets for research purposes. This network will enable data from a diverse and rich range of health datasets to be linked across jurisdictions and sectors in order to facilitate nationally and internationally significant population level research, to improve health and wellbeing and enhance the effectiveness and efficiency of health services in Australia.

### **2.3. Centre for Data Linkage**

The Centre for Data Linkage (CDL) is a component of the national PHRN project and was established within Curtin University as part of the NCRIS initiative. The focus was to develop and implement secure, state-of-the-art national infrastructure to enable cross-jurisdictional data linkage for research.

Australia's federated government system means that various datasets are gathered at different tiers of administration and that different jurisdictions are responsible for different data collections. To realise the full potential of these data resources, it is necessary to link data between these jurisdictions to ensure complete population coverage. Therefore, the PHRN initiative is unique, providing cross-jurisdictional linkage i.e. linkage of data across nine different legal jurisdictions (seven States, two Territories and the Commonwealth). Despite significant investment in data linkage in Canada and the UK, no group has attempted to systematically link data across countries, states or provinces.

### **2.4. Designing Secure Linkage Infrastructure**

Production linkage on a national scale requires the ability to provide levels of availability, service, access and performance that are potentially a magnitude larger than existing state

or territory based data linkage facilities. In terms of an operating environment, the objectives for the national infrastructure were:

- 1) The ability to provide secure linkage systems and services to internal and external stakeholders, with adequate levels of availability;
- 2) An environment that is auditable and certifiable against the PHRN Information Governance Framework, and other relevant industry standards;
- 3) Cost effective, low maintenance environment that can draw on shared services within a provider's managed environment (e.g. software updates, licencing, networks, directories, security technologies).

The final model builds on the foundation provided in the Briefing Paper: Population Health Research Network Centre for Data Linkage Proposed Model (distributed to stakeholders in December 2008) and takes into account comments and feedback received from other Network participants.

## **2.5. Cross-Jurisdictional Operational Model**

The infrastructure is designed to provide a platform for undertaking large, national linkage projects while meeting the requirement to provide a secure and controlled environment for working with sensitive data. The Operational Model concerns the conduct of the linkage units core purposes of facilitation of linkage between jurisdictional datasets and is designed to maintain security, privacy and to scale as dataset size and demand for national linked data increases.

The infrastructure is designed to support the following core functions:

- Provision of demographic information from data custodians to the CDL;
- Linkage of this data to create project specific national keys;
- Supply of keys back to the various data custodians in each jurisdiction; and
- The extraction and transfer of the necessary clinical data from the jurisdictional custodians to the researcher.

These cross-jurisdictional data flows are represented in Figure 1.

### **Demographic information**

In order to generate the National Linkage Map, the linkage facility requires individually identifiable demographic information about the individuals in the participating administrative

databases. The National Linkage Map is central to the linkage model and consists of 'pointers' to records in participating data collections. Although the creation of the map requires access to individually identifiable demographic information, these data items are not contained in the National Linkage Map.

### **Experimental design – Separation of data linkage and analysis**

The PHRN was very aware of the sensitivities associated with maintaining databases containing individually identifiable information; therefore the final model incorporates the "two stage", privacy preserving linkage model used in Western Australia and New South Wales [52] to link demographic data from multiple jurisdictions. This approach distinguishes between (1) activities used to link data and (2) activities used to generate linked re-identifiable datasets for approved research projects. In this model, health records comprise two components:

1. the 'demographic' component, comprising individually identifiable information about a person such as their name, address and date of birth; and
2. the 'health' element, containing the sensitive information about a person's health such as the procedure they had during a hospital stay or details of cancer type if they have been diagnosed with cancer.

The model is implemented in a number of ways, including separate storage of demographic data away from 'pointer' information which is stored in the National Linkage Map. In addition, the model also separates linkage functions i.e. staff members who deal with the demographic data (that is, the linkage team) are different from administrative staff and any other persons outside of the CDL who access the health data (that is, the data custodian and the relevant project investigators).

In Stage 1, the linkage team (the CDL) are supplied with demographic data from participating datasets. The CDL uses this information to generate the National Linkage Map. The Map is a set of high quality linkage keys that can identify the same person within and between datasets. Once established, the Map is kept separate from the demographic data used to create it. The Map will be maintained for the duration of the project. For privacy purposes, the linkers will not have access to details of the health component of a record - this remains under the full control of the relevant dataset's custodian.



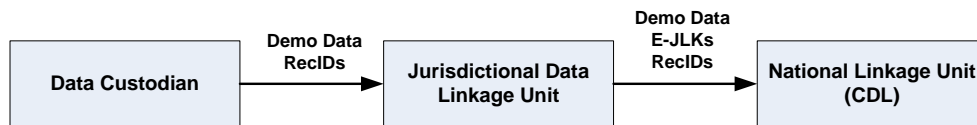
In Stage 2, linkable health data is released by the original data custodian (not the CDL) for approved research projects. Where a researcher has approval to access a linked dataset, the data custodian for each dataset prepares a file containing health information items from a designated set of records.

### Project Specific Keys

Another essential element of CDL operations is the generation of project specific keys. These are generated by the CDL for each approved research project. The project-specific keys are supplied to relevant data custodians and used by them to extract clinical data for researchers. Each research project is allocated a different set of project keys so that researchers working on various projects who each receive separate clinical data cannot later combine their datasets. Importantly, the CDL never releases information from the National Linkage Map. The Map and its contents remain as a "master copy" which is encrypted and kept under the control of the CDL.

**Figure 1: Cross-jurisdictional data flows**

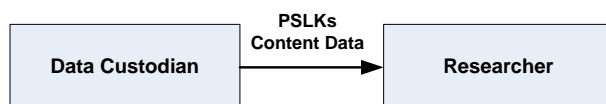
Step 1: Flow of Demographic data for linkage (once per dataset)



Step 2: Issue of Project-Specific Linkage Keys (once per project)



Step 3: Flow of content data for research (once per project)



**Key**

- Demo Data: Demographic Data
- RecIDs: Record ID numbers
- E-JLKs: Encrypted jurisdictional linkage keys
- PSLKs: Project specific linkage keys
- Content Data: Content or clinical data used by researchers

## 2.6. Designing National linkage infrastructure

The national linkage system is designed to undertake linkage across event datasets, based on probabilistic matching of demographic information, group these events, and process mapping requests from jurisdictional data providers by supplying encrypted project-specific identifiers to release for researchers [70].

Objectives of the development process were to:

- Determine the features to be implemented in the 'baseline' production system;
- Design and develop working software which will realise this initial set of features; and
- Address the unique security and volume requirements of a data linkage system.

To future proof the secure data linkage facility, it was important to create infrastructure that can scale as dataset size and demand for national linked data increases. The process included preparation of detailed Technical Specifications, Feature List and Technical Feasibility for the development.

## 2.7. Functional requirements

Based on the agreed operational model for national and cross jurisdictional linkage, which requires efficient and accurate processes, the high-level system requirements necessitated a system which was reliable, easy to maintain and operate, with auditing capabilities. From the software evaluation, it was clear that almost all of the system could not provide a robust enterprise-grade platform which could easily scale to the data sizes anticipated for national and cross jurisdictional linkage in Australia.

To ensure a national linkage system that was 'fit for purpose'; software was developed to include linkage and management capabilities. The system was designed using a component approach which focused on system integration, interoperability and expansion capabilities to ensure future flexibility. The 'baseline' development criteria included the following requirements:

*Secure and auditable* – to ensure transparency of operation, the system had to be secure and provide an audit trail for all system actions. Security was implemented in a role-based access control model. This method regulates access to the system based on the roles of individual users. The system roles and their implementation have been defined as part of the system architecture and are managed using standard operating procedures. User roles can be created, changed, or withdrawn as the needs of the service change, without individually updating the privileges for every user.

*Enduring and project linkage* – To maximise flexibility the system manages a range of projects from a simple ‘one-off’ project with a short life span through to enduring longitudinal datasets constructed through the linking of records from successive time periods. The ability to manage both types of projects ensures flexibility and versatility in linkage operations.

*Data volume* – all system components (load, linkage, data management and output) had to have the capacity to handle large data volumes. This was crucial to managing projects which involve tens of millions of records and billions of matching transactions. The system had to have the capacity to scale as projects get bigger to avoid redundancy.

*Project management* – The system is designed to manage multiple projects without an overhead to performance. The user interface provides operators with the ability to create and manage projects (linkage and extraction), custodians and data. The system can manage multiple large projects without substantial or complex operator involvement.

*Link management* – unlike most linkage systems, the software manages changes in data and links over time. This means that the software can automatically process amended and deleted records as well as adding new records. Unlike other designs, the system stores and processes all linkage transactions at the matching pair level. This allows the system to automatically detect and manage change to the linkage map as data is added (including new records and amendments to records). It also supports ‘any point in time’ referencing at the group or map level allowing operators to recreate the linkage structure for any records at any (previous) point in time. This functionality was not available in any of the software evaluated.

## **2.8. Development of an automated production linkage system**

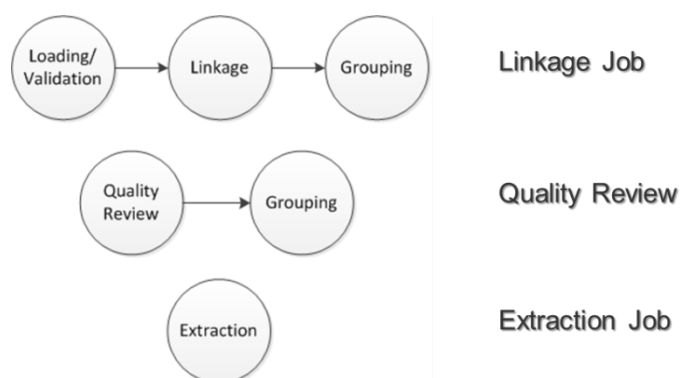
Development of the baseline system was built around the research and prototyping undertaken as part of the environmental scan and software evaluation. The system was developed over a series of construction and release iterations and has undergone considerable User Acceptance Testing. As the software was developed, the priorities of selected features were reviewed and reassessed to ensure that software development aligned to the business priorities of the CDL. Future development and enhancement will build on this initial version.

The final architecture uses enterprise grade databases (Oracle or SQL Server) to store all records and transactions. The system is managed through a web interface which allows role

based access to linkage functions and features. The system automates linkage functions that traditionally require manual intervention (including linkage and quality processes).

Administrative and monitoring functionality allows operators to manage linkage projects and data from different data providers. The linkage process within a project can be specified in three 'core' procedures: linkage, quality and extraction. These can be monitored through the user interface, and built-in audit trails track operations on all processes, data (records or transactions) and data custodian details.

**Figure 2: Linkage processes**



*Linkage* – The matching strategy, defined by the linkage operator, is automatically managed through three stages by the system. First, the data to be linked is loaded into the system; this process includes validation and standardisation of records prior to matching (rejected records are removed and a report provides reasons for exclusion). The matching configuration for each linkage can be designed by the operator (allowing changes to blocking, comparisons and scoring). Matching is carried out by the linkage engine which produces all pair comparisons based on the match configuration. The system generates dynamic groups from the pairs which are managed over time by the system.

*Quality* – Traditionally quality intervention had been a manual process which can only be implemented after all linkage procedures have been completed. This automation of the quality transactions allows the system to manage quality interventions and to highlight any inconsistencies (suspicious groups).

*Extraction* – Project extractions are controlled by the system and produce a project specific linkage map for the researcher. The project keys generated in the map are only relevant to

an individual linkage project (even if the dataset appears in another linkage). This ensures no crossover between projects or project teams.

## **2.9. Performance evaluation of linkage engines**

A performance evaluation of available data linkage software packages was undertaken to support the decision making process in regard to the choice of linkage software (and to provide further information to other PHRN nodes).

This involved identifying, procuring and installing a shortlist of suitable linkage software. A performance and linkage quality review of each program or package was completed. The transparent and transportable performance evaluation framework was developed specifically for this project but has been shared with many researchers to enable assessment of linkage systems. The evaluation framework uses synthetically generated “standard” datasets and a standard linkage approach (strategy) in the performance review of each package. Where possible, the packages were evaluated on a standard hardware configuration so that performance aspects could be fairly compared. A review of the functionality and features of each software package was also undertaken.

## **2.10. Linkage management and matching engine**

It was clear from the evaluation that very few of the linkage packages would provide the performance and functionality required to both manage and transform datasets as linkage demand, complexity and size grow. The research programme required a system to be developed which could provide scalable, fast, efficient, accurate and cost effective linkage.

To build capacity and to ensure timely research data for linkage projects, the research programme created new enterprise grade linkage infrastructure to provide a platform to run large linkage projects. The resulting linkage system includes the functions and features identified during the evaluation process and incorporates the theoretical framework described in this thesis. It has also been designed to be both linkage engine and database agnostic allowing future changes to these components.

The linkage system currently uses CUsomisable Probabilistic Linkage Engine (CUPLE) as a matching engine. This linkage engine was developed at Curtin and was designed to provide scalable matching performance and to harness developments around deterministic, probabilistic and machine learning linkage. It provides a configurable framework allowing operators to customise blocking options, comparison routines and scoring for each linkage project. CUPLE is multi-threaded to improve performance (and provide hardware scalability). The linkage engine produces matching pairs based on the matching configuration.

A separate grouping or clustering process, which is managed through linkage system, then amalgamates these record-pairs into groups to identify the full set of records belonging to the same individual. The 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.

## **2.11. Governance**

The challenge for the CDL, and other organisations with the need for managing biomedical data privacy, is to translate information governance frameworks and standards down to a set of rules, concepts and designs that can be implemented as cost effective technical solutions [71]. The model development involved working with IT departments (or outsource providers) whose priority and expertise is on supporting corporate systems (e.g. Finance and HR), not necessarily dealing with the specialised needs of linkage researchers and analysts [72].

As part of the design process, the research project developed a set of guidance infrastructure architectures or 'Design and Implementation Guidelines' for a secure Research Computing Environment to host the CDL National Linkage System and supporting applications [73, 74]. The design allows the CDL to store and use the data provided for each linkage project in a highly secure environment. This includes physical security features (such as key card access to the CDL office, additional card access for entry into the secure computer room and a safe to store protected information in physical form e.g. DVD) as well as technical security measures (such as computers requiring password login, automatic screen locking and monitoring of login attempts) and data security (e.g. the use of encryption to store information).

## **2.12. Ethics approval**

Approval from the Curtin HREC has been obtained to establish the core operations of the national linkage system, that is, the capacity to receive demographic data to generate the National Linkage Map. In addition to this approval, other state-based HREC approval was required to allow construction of national linkage maps for specific projects so that state and territory data providers can release their demographic data for linkage. These approvals do not overlap with the oversight jurisdiction of the Curtin HREC. In addition, the creation of the National Linkage Map for each project requires individually identifiable demographic information sourced from databases owned by participating data custodians.

## 2.13. Conclusion

This chapter addresses the second aim identified in the thesis to develop an operational model and robust governance framework for national linkage infrastructure. The greatest risks associated with the model for national linkage relate to the possible breach of privacy through the disclosure of personally identifying health information. The likelihood of this outcome is low, as a number of strategies have been implemented to reduce the risk. These strategies include highly physical, technical and procedural security, and strong information governance.

Risks have also been minimised through the use of a best practice protocol [52] which restricts the release of identifiable data to limited data items. The data items include demographic data (including name, sex, data of birth or address) but exclude any health information. The demographic data fields are necessary to undertake accurate data matching across (and within) different datasets. This model, and associated data flows, is described in detail in the paper entitled 'Data linkage infrastructure for cross-jurisdictional health-related research in Australia'.

This CDL cross jurisdictional model presented in the paper provides significant protection of patient privacy compared to methods where both named information and personal health information are disclosed. Under this model, only specific data items that are needed to produce high-quality data linkage results will be requested and used by the linkage team. No personal clinical information is released for data linkage. Dedicated linkage facilities also lead to increases in the accuracy and reliability of linkage results and improvements in the value and quality of routinely collected data.

It has been shown in Western Australia that record linkage has significantly reduced the invasion of privacy associated with use of confidential health information. A small linkage group that has access to personal details but not to clinical information can replace the personal details with unique project identifiers and release these to researchers, who can thus link multiple sources of clinical information without ever having access to personal details. Based on evidence from the WA Data Linkage Branch (the state-level data linkage unit based out of the WA Health Department), the proportion of research projects requiring named data fell from 94% in 1994 to 36% by 2003.

There is considerable evidence that benefits to the community and individuals in improved quality of health care achieved through data linkage methods considerably outweigh the risks potentially arising through breach of privacy and confidentiality [51, 68].

The chapter also addresses the third aim to investigate, design and build dynamic national data matching infrastructure for Australia. This included an evaluation of linkage methods and software to assess the strengths and weaknesses of available systems using standard performance metrics. The paper by Ferrante and Boyd, describes the transparent and transportable evaluation method which has since been used by international colleagues to benchmark a range of matching products.

The final system design was influenced by functionality in existing linkage infrastructure nationally (Western Australia and New South Wales) and internationally (Scotland). Practical aspects of providing linkage 'as a service' are described in 'Technical challenges of providing record linkage services for research'. The paper outlines a number of linkage scenarios along with associated operational requirements to support research. The core components highlighted include data management, process automation and the ability to maintain the linkage map over time. These are key concepts used in the technical design of the national linkage system.



## 2.13. Published Manuscript(s)

**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 services research (2012)**



CORRESPONDENCE

Open Access

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

James H Boyd<sup>1\*</sup>, Anna M Ferrante<sup>1</sup>, Christine M O'Keefe<sup>2</sup>, Alfred J Bass<sup>3</sup>, Sean M Randall<sup>1</sup> and James B Semmens<sup>1</sup>

## 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

<sup>1</sup>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:

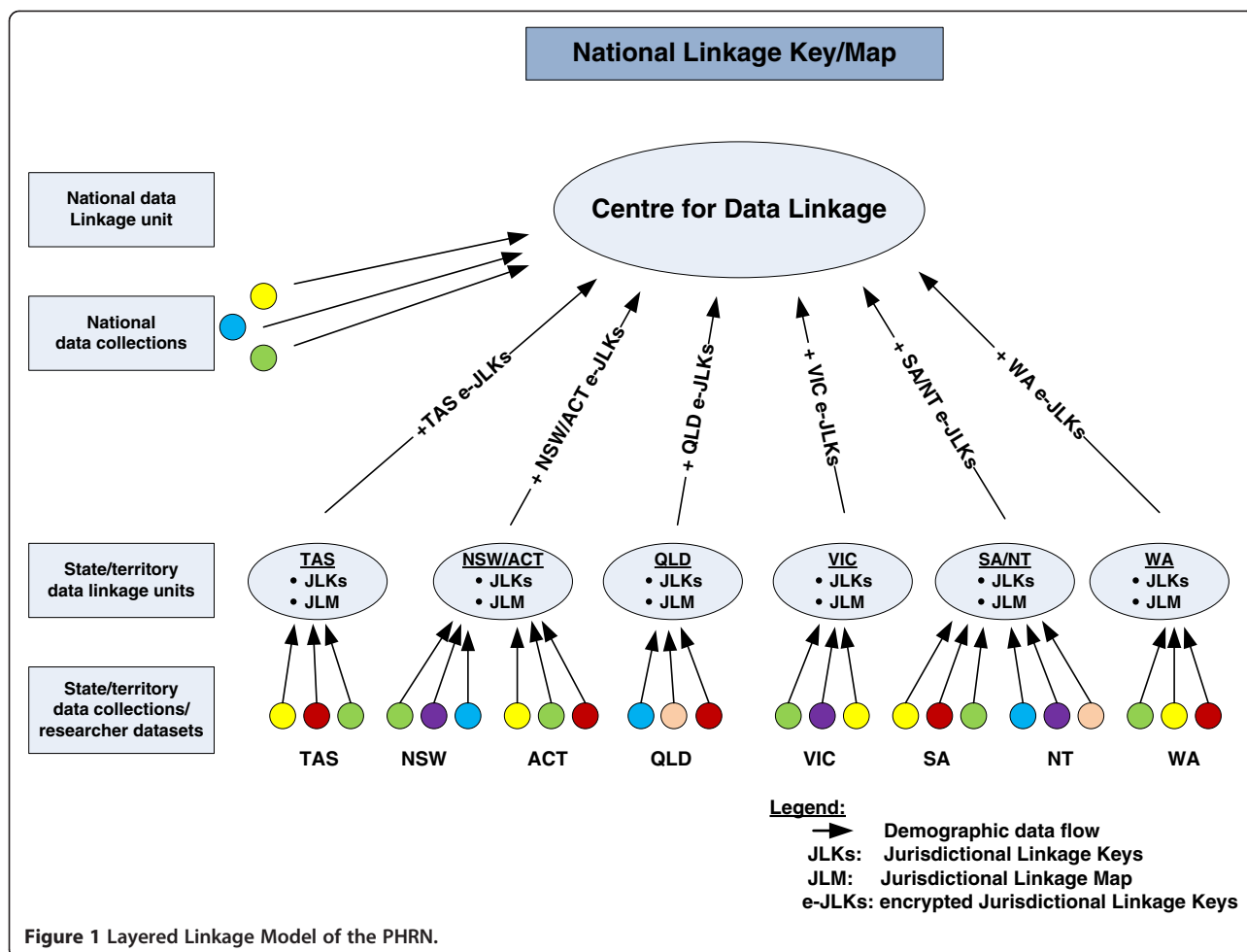
- i) Development of a cross-jurisdictional operational model
- ii) Specification and implementation of a secure IT environment including linkage software; and
- iii) Development and adoption of strong governance arrangements

#### **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 influence 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



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

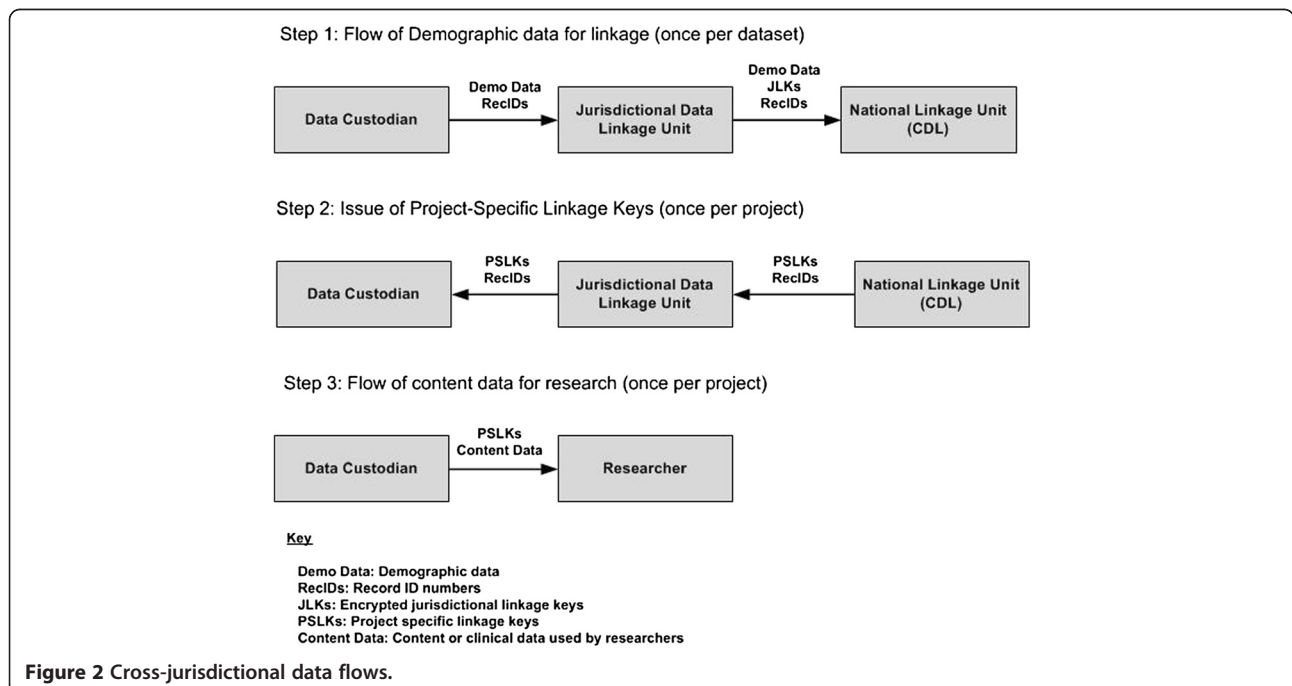
dataset to create 'linkage keys'. Clinical or service information is not needed by the CDL and is not provided to it and the researcher receives only that part of the record that they have approval to see (without any demographic or identifying information).

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



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.

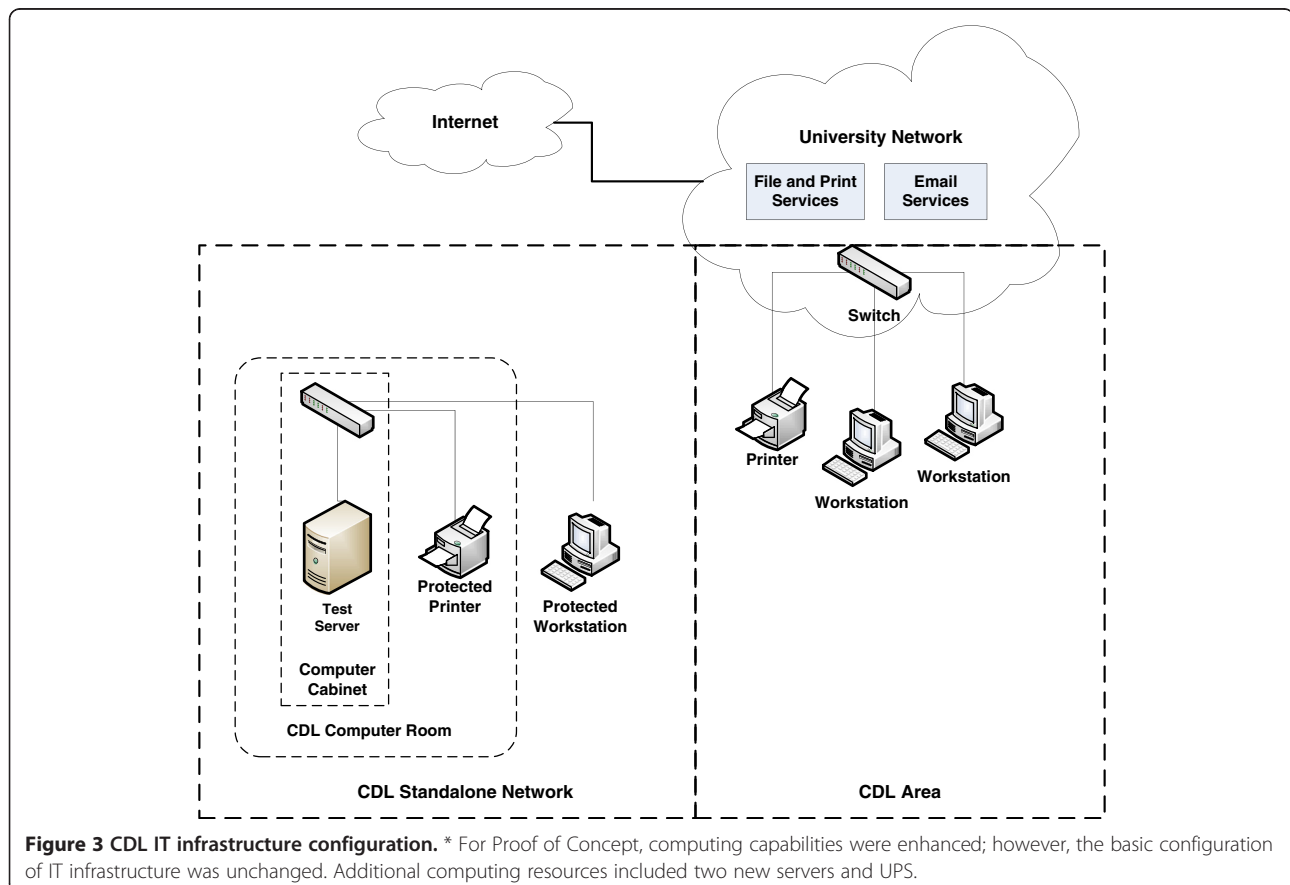
Among other things, the audit report provided an assessment of how the infrastructure was configured and

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



strict reporting regime; a Privacy framework; an Information Governance framework; rigorous approvals processes for each research project; binding agreements related to data release, data 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

<sup>1</sup>Curtin University, Perth, Western Australia. <sup>2</sup>CSIRO Mathematics, Informatics and Statistics, Canberra, ACT, Australia. <sup>3</sup>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, Kiewer 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 Population-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, Verplanck 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.

**Boyd JH**, Randall SM, Ferrante AM, Bauer JK, Brown AP, Semmens JB. *Technical challenges of providing record linkage services for research*. BMC Medical Informatics and Decision Making (2014)



**CORRESPONDENCE**

**Open Access**

# Technical challenges of providing record linkage services for research

James H Boyd<sup>1\*</sup>, Sean M Randall<sup>1</sup>, Anna M Ferrante<sup>1</sup>, Jacqueline K Bauer<sup>1</sup>, Adrian P Brown<sup>2</sup>  
and James B Semmens<sup>1</sup>

## 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

<sup>1</sup>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

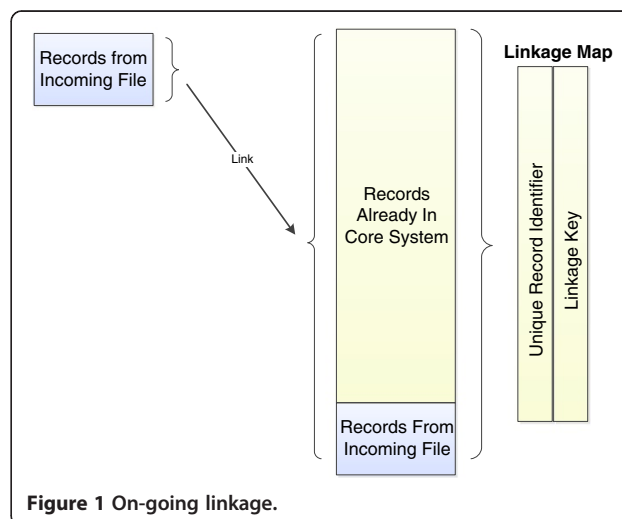


Figure 1 On-going linkage.

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

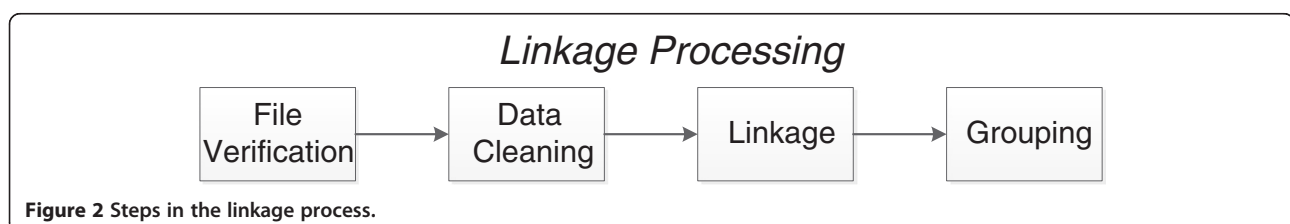


Figure 2 Steps in the linkage process.



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<sup>a</sup>.

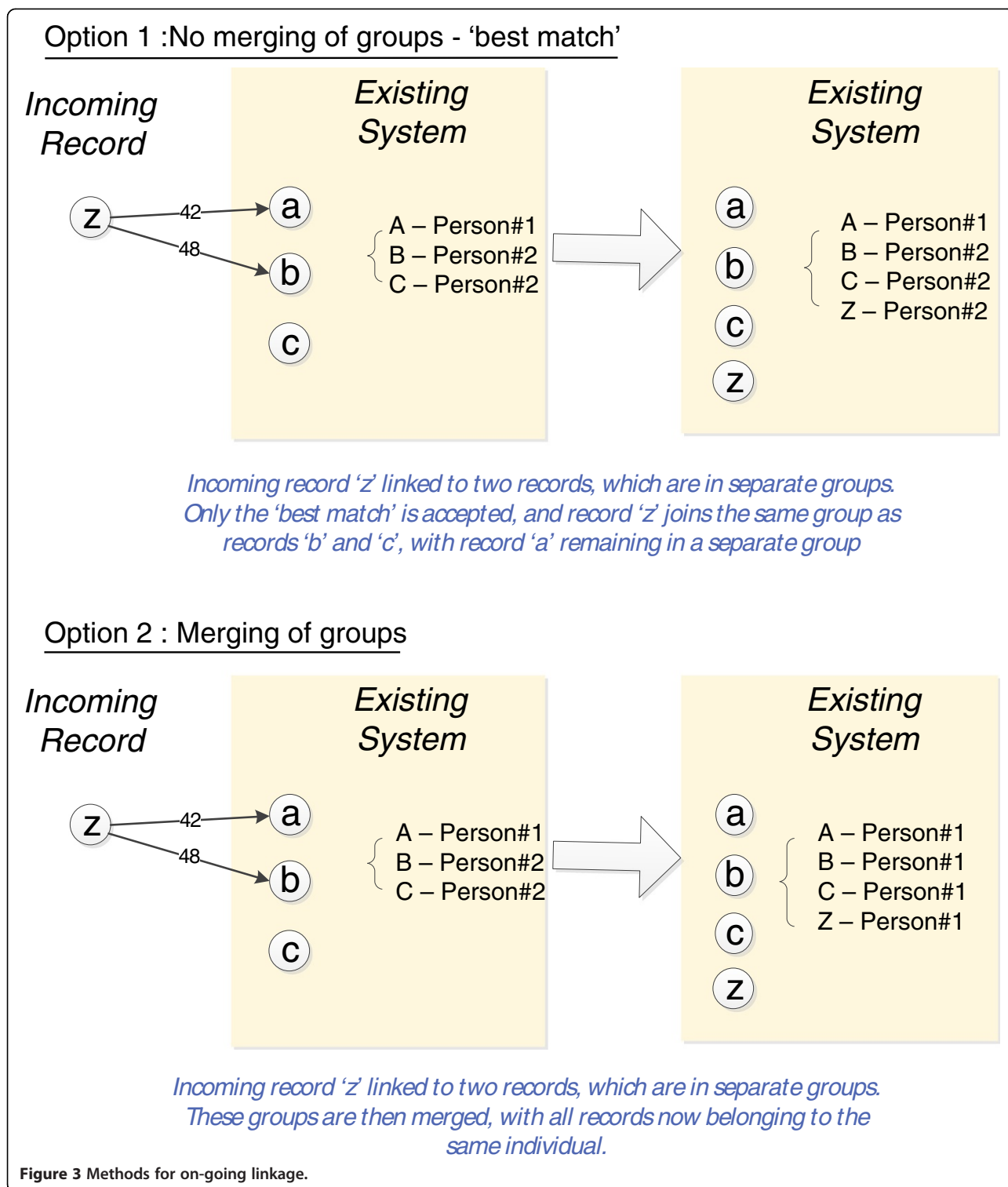
An alternate approach is to allow the merging of groups to occur. This method does not rely on the existence of a high quality reference dataset (spine). For this reason this method may be useful in a much greater range of circumstances.

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.

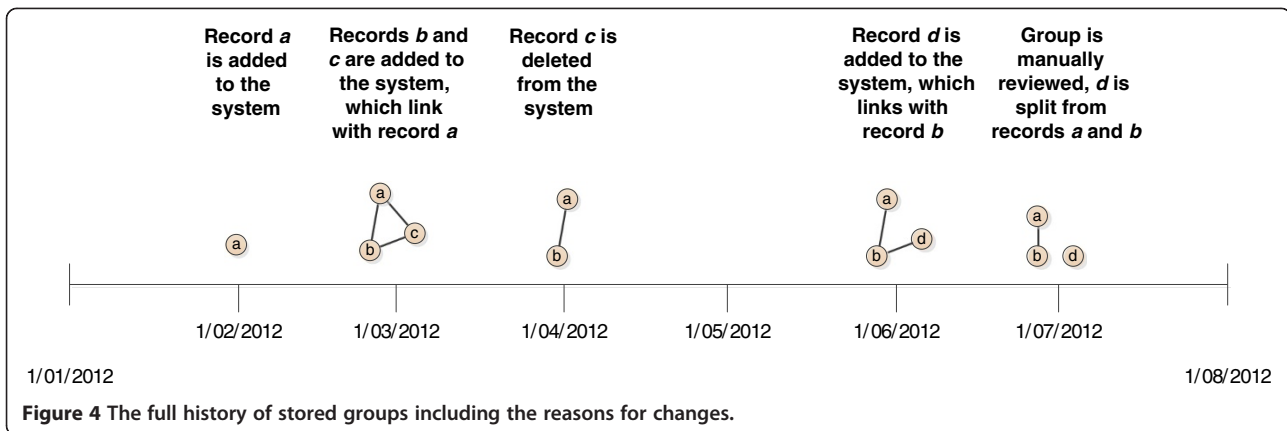


**Managing deleted, amended and 'open' records**

**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



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).

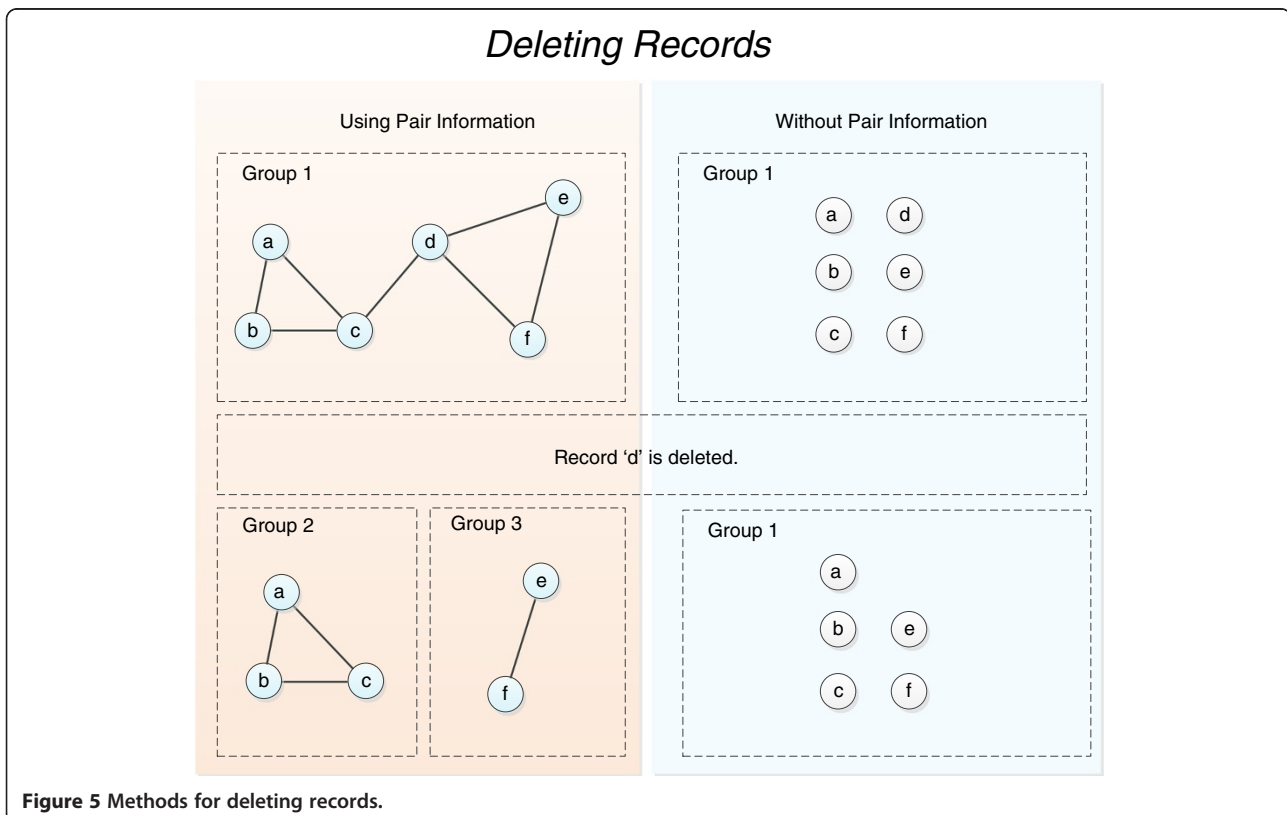
**Amended records**

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



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	<ul style="list-style-type: none"> <li>- Link to most recent record in group vs. link to all records</li> <li>- Best-link matching vs. merging groups</li> </ul>
Linkage automation	<ul style="list-style-type: none"> <li>- Spectrum from fully automated to only the grouping process automated</li> </ul>
Links stored	<ul style="list-style-type: none"> <li>- No history stored</li> <li>- Snapshots stored</li> <li>- Full history stored within linkage map</li> </ul>
Handling different linkage scenarios	<ul style="list-style-type: none"> <li>- Only on-going linkage</li> <li>- Manual processes for project based linkage</li> <li>- Access to on-going linkage system used for project based linkage</li> <li>- Build system which can handle multiple scenarios</li> </ul>
Amended and deleted records	<ul style="list-style-type: none"> <li>- No handling of amended and deleted records</li> <li>- Amended records: Changing personal identifiers only vs deleting and re-linking</li> <li>- Deleted records: Simple removal, or using pair information to reconstitute groups</li> </ul>

## Endnote

<sup>a</sup>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

<sup>1</sup>Centre for Data Linkage, Curtin University, Perth, Western Australia.

<sup>2</sup>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, Verplanck 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](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 Computing 1996.* 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.



Ferrante AM and **Boyd JH**. *A transparent and transportable methodology for evaluating Data Linkage software*. Journal of Biomedical Informatics (2012)





## Chapter 3

---

### Privacy and data linkage

*“Historically, privacy was almost implicit, because it was hard to find and gather information. But in the digital world, whether it's digital cameras or satellites or just what you click on, we need to have more explicit rules - not just for governments but for private companies”*

*Bill Gates*

#### **Book Chapter(s):**

**Boyd JH**, Randall SM, Ferrante AM. *Application of privacy preserving techniques in operational record linkage centres*. Medical Data Privacy Handbook. Springer International Publishing, 2015. 267-287.

#### **Published Manuscript(s):**

Randall SM, Ferrante AM, **Boyd JH**, Bauer JK, Semmens JB. *Privacy-preserving record linkage on large real world datasets*. Journal of Biomedical Informatics 2014; DOI: 10.1016/j.jbi.2013.12.003

#### **International Conference presentation(s):**

Randall SM, **Boyd JH**, Ferrante AM, Bauer J, Gillies M, Semmens JB. *Privacy preserving record linkage on large real world datasets*. Conference: International Health Data Linkage Conference, Vancouver, Canada, April 2014.



### 3.1. Privacy challenges in Record Linkage

Administrative data collections are highly confidential, often containing sensitive personal information that is protected by law. Australian privacy laws permit some level of disclosure of personal information by authorities for human research (*Commonwealth Privacy Act 1988* s95) [75]. Critical in the decision making required to release personal data is striking a balance between the use of personal records for public good (research) and ensuring the privacy of individuals [76].

Data custodians and researchers have developed data access and usage models that comply with information privacy laws and provide important safeguards to privacy. Dedicated record linkage centres with secure environments and specialised linkage personnel have been established to support health research as well as for routine reporting, policy development and service design by government [77]. These record linkage centres have developed and implemented various strategies to minimise the risk to privacy posed by their operations [78]. These strategies and techniques can be grouped into three broad categories:

- *Data governance*: policies and procedures around data access, data transfer and IT security;
- *Operational models and data flows*: organising and controlling the movement of information to minimise information disclosure risk; and
- *Privacy preserving linkage techniques*: developing and testing methods of accurate data matching without requiring personally identifying information.

### 3.2. Data governance

To maximise privacy and maintain confidentiality, integrity and availability of the personal information used for record linkage, it is necessary to develop a governance framework which:

- Provides a consistent approach to good information governance which demonstrates that record linkage centres and stakeholders understand, prioritise and manage risks associated with data transmission, access, use, storage and disposal;
- Recognises the need for flexibility given the organisational environments and different business requirements of record linkage centres and stakeholders; and
- Takes into account existing information management and security policies, practices, processes and infrastructure of record linkage centres and stakeholders.

In Australia, data custodians, researchers and record linkage centres have worked together to develop data access and usage models that comply with information privacy laws and provide sufficient guards to privacy (e.g. Integrating Authority principles [79], CDL model). Record linkage centres, in particular, have implemented best practice data governance frameworks to minimise the risk to privacy posed by their operations [41, 80-84].

### **3.3. Operating model and data flows**

A number of possible operational models can be employed to ensure linkage is carried out efficiently and securely. Each of these operational models has strengths and weaknesses and their applicability or suitability will depend on institutional setting and governance arrangements. An outline of some of the more conventional models is provided.

#### **3.3.1. Centralised model**

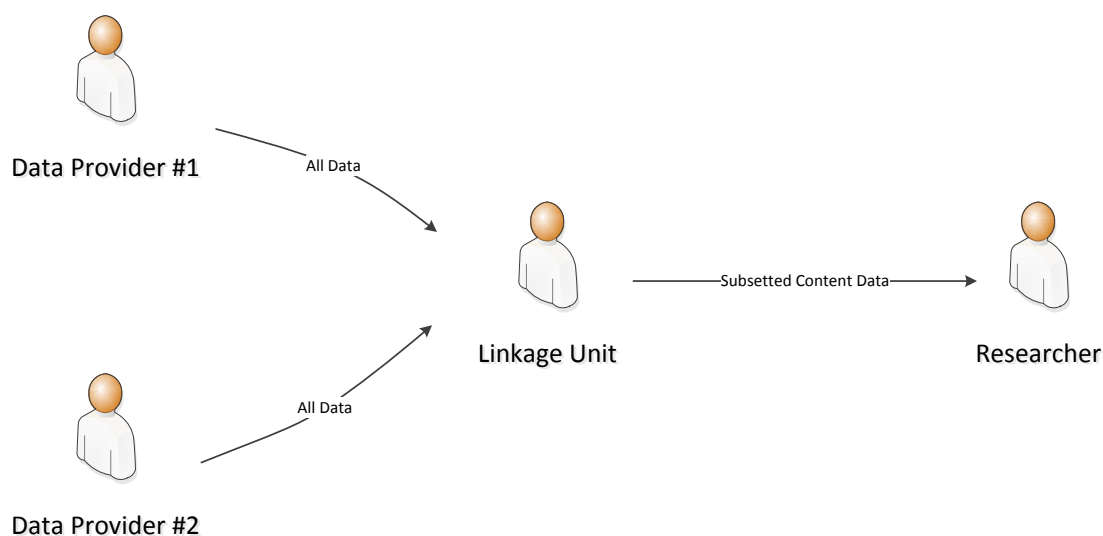
In a centralised model, data from providers are supplied to a central linkage facility for matching. This includes both personally identifying information as well as clinical content data. Using personally identifying data the central linking facility constructs a linkage map with a set of linkage keys that identify the same individual within and across multiple datasets. The clinical data can then be extracted by the central unit for the researcher once they have received approval.

In this model, there is minimal separation between the linkage unit and client services teams.

The main features of this model are:

- Centralised data collection; and
- Centralised linkage.

**Figure 3: A centralised model: Data providers give their full datasets to the linkage unit, who link and then pass on the content data required for research to the researcher**



**Advantages:**

- Centralised systems are easier to manage and maintain than distributed ones;
- Less complicated data flows mean potentially quicker turnaround;
- Linkage quality may be improved by access to clinical data.

**Disadvantages:**

- Requires release of name-identifying and content information to the same organisation - a potential privacy risk;
- Data providers may not be comfortable with this additional privacy risk and opt-out of providing data, reducing potential demand for this service, as well as overall linkage quality.

### **3.3.2. Separated model, with centralised clinical data**

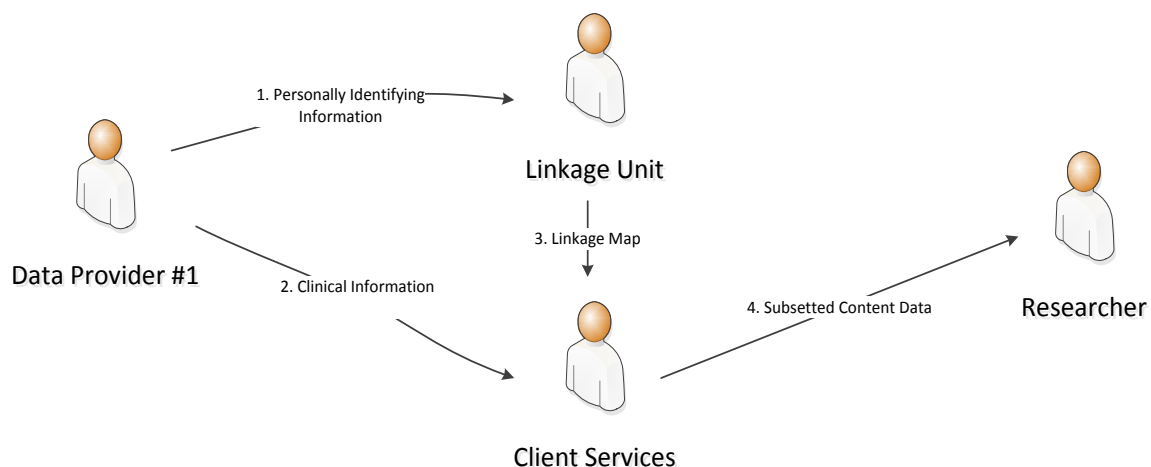
To minimise privacy and confidentiality risks, the personally identifying information supplied to operational linkage units is often separated from clinical information prior to release by data custodians. In addition to the separation principle, these linkage units have adopted strong policies and secure procedures for the handling, use and disclosure of personally identifiable information.

By separating clinical data from personal identifiers during the linkage process, the risk of revealing sensitive information about individuals is dramatically reduced. Nevertheless, some residual risk to privacy remains. Ideally, data custodians seek a zero-risk method of

providing accurately linked research data without the need to disclose any identifying information to linkage units. The separation principle is widely used in Australia. [52, 79, 85]

In this model, only the personally identifying information required for linkage is supplied to the linkage unit. The clinical data is passed to the client services team. It is important here that the linkage team is separate from the client services team in the case where they are both part of the same organisation. Once the linkage is performed on the personally identifying information, the linkage map is passed to the client services team who join this to clinical information to create datasets for research and analysis. The client services team then extracts information for the researcher.

**Figure 4: The data provider splits their data, sending personal identifiers to the linkage units, and clinical content to the client services team. The linkage unit then provides the linkage map to the client services team who join it to content data to create datasets for research and analysis.**



**Advantages:**

- Increased privacy by adhering to the separation principle. Only the data provider has both the clinical data and the personally identifying information for each data collection;
- This model has centralised clinical information and centralised person identifiers (albeit not in the same place) - centralised systems are easier to manage and maintain than distributed ones.

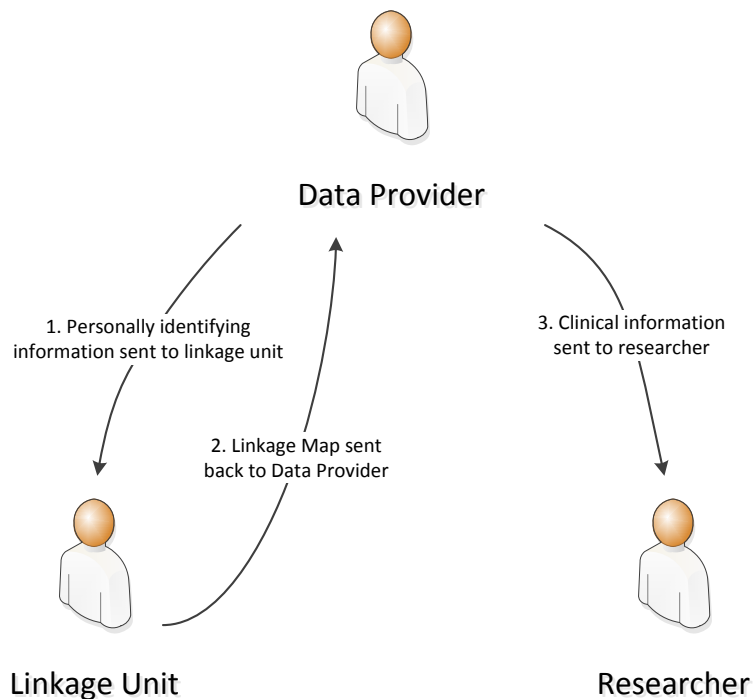
*Disadvantages:*

- In the situation where the linkage unit and the unit managing the clinical information are part of the same organisation, there may be perceived risk of individuals having access to both sides of data;
- Linkage quality can't use clinical data for quality assurance purposes.

### 3.3.3. Separated model, with no centralised data repository

The linkage unit receives only the personally identifying information required for research. Instead of handing the created linkage map to the 'client services' team who hold the clinical information, this clinical information is not centralised and is held separately by each of the data providers. The linkage unit returns each required portion of the linkage map to the separate data providers. Each data provider then extracts the data for the researcher. Under this model, the researcher is responsible for creating the final, linked research datasets.

**Figure 5: In the absence of a repository of clinical data, this is supplied to the researcher by the data provider**



*Advantages:*

- Increased privacy, with personally identifying information only given to the linkage unit, and clinical information only given to the researcher.

*Disadvantages:*

- This model requires data providers to play an active role in linkage operations. Many data providers may not see this role as part of their core service, and may not be able to provide this functionality;
- With more complex data flows, there are more opportunities for mistakes, and turnaround times may be longer;
- Depending on the research question, researchers may need to be provided with a larger amount of records than is entirely necessary for their study. This is because the sub-setting of records based on clinical characteristics typically does not occur in a decentralised model;
- Researchers will receive separate data files from each data provider, which they will have to amalgamate.

### **3.3.4. Operational models involving multiple linkage units**

In the situation where there are multiple linkage units operating, there are several possible ways in which the units may work together.

One option is for linkage units to operate entirely independently from each other. Depending on project requirements, each unit would receive data from data providers (possibly, from the same data providers) and link this independently.

An alternative option is for linkage units to work cooperatively on projects, receiving and amalgamating linkage maps from each other to produce a single linkage map. By using another linkage unit's linkage map, significant quality work which they may have invested in can be leveraged. Personally identifying information from data providers could be received through these linkage units, leveraging off the existing relationships these units have with their data providers. This model has been adopted in Australia to enable cross-jurisdictional data linkage [85].

## **3.4. Privacy Preserving Record Linkage Methods**

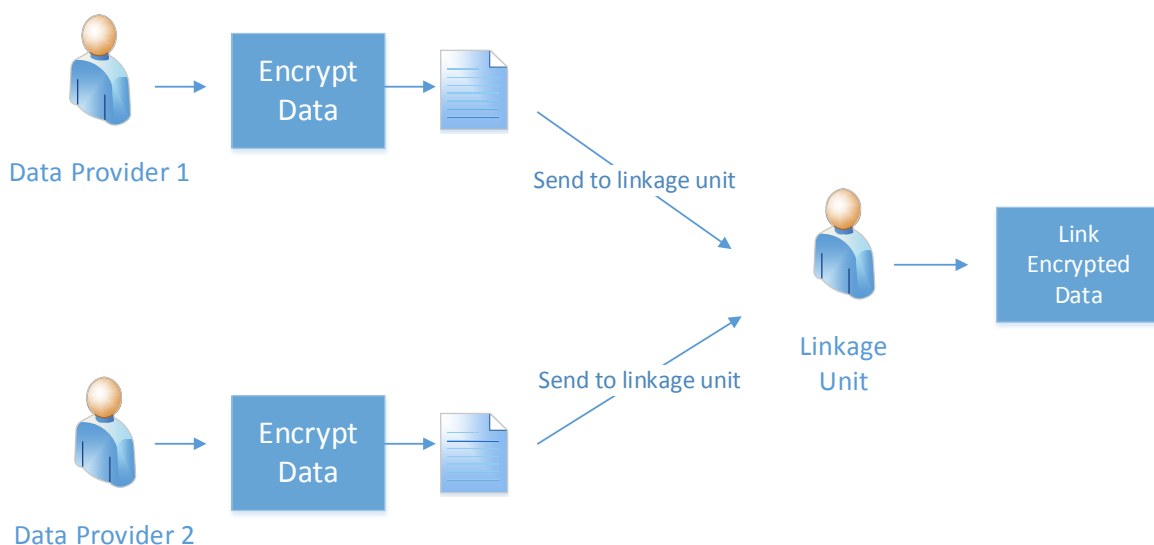
One way of further reducing the privacy risks associated with record linkage is through privacy preserving record linkage (PPRL) techniques [86]. PPRL methods undertake data matching without requiring personally identifying information. Instead, PPRL algorithms operate on information which is hashed or encrypted by custodians before release to third parties. Having been transformed into a permanent, non-identifiable or "privacy-preserved" state, the data is then supplied to record linkage centres and used in probabilistic record linkage.



Privacy preserving techniques generally adopt the same security model as unencrypted linkage; however, there may be differences in the particular privacy algorithm used. Nearly all privacy preserving protocols take an ‘honest-but-curious’ threat model [87], whereby parties are expected to try to carry out the protocol correctly, but will also try and find out as much as they can from any data received.

In these PPRL protocols, data is encrypted in a way which allows linkage to be carried out, without personal identifying information being disclosed. Data is encrypted by the data custodian (using specially provided software) before being sent to the linkage unit, who carries out the linkage (see Figure 6). The encryption is irreversible and different outputs can be generated for projects (project-specific encrypted data).

**Figure 6: Privacy preserving record linkage process**



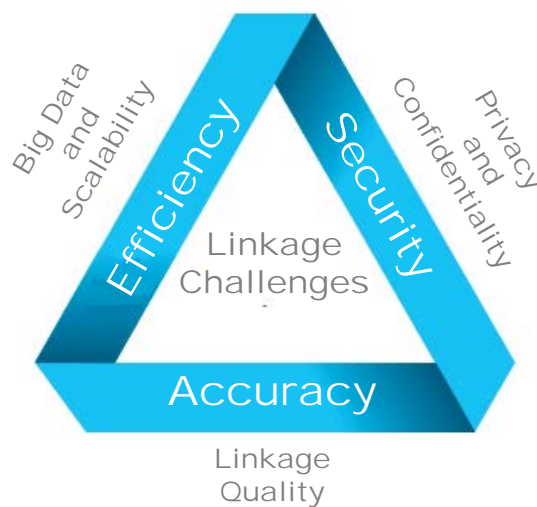
A variety of privacy preserving techniques exist. These include: random value, secure multi-party computation (SMC), bloom filter, embedded space, generalisation (k-anonymity, binning), grams projection, multi-bit tree and phonetic encoding [88-94]. Privacy preserving protocols range in the comparison techniques applied, from those carrying out an exact match of entire records to those employing string similarity measures on individual fields. Protocols utilising more fine-grained techniques in determining similarity will yield higher linkage quality [95].

The standard linkage processes (transformation, blocking, matching) remain at the core of PPRL. However, constraints within each of these steps need to be carefully managed to avoid issues (i.e. degraded content information to increase privacy may compromise matching accuracy or additional processing to improve linkage quality could be computationally expensive leading to excessive processing times). A challenge in the adoption of privacy preserving methods is to balance these constraints to achieve secure and efficient linkage with a high matching accuracy when applied to real world contexts.

### 3.5. Balancing the privacy constraints

Few of the PPRL algorithms have been practically evaluated on the basis of their suitability for use in operational record linkage settings [96]. For privacy preserving record linkage techniques to be usable in an operational context, they must be not only secure but also accurate and efficient (Figure 7) [97].

**Figure 7: Privacy Preserving Constraints**



**Security:** Although PPRL techniques significantly reduce privacy risks, some (small) residual risk remains. Recent research suggests that, under certain circumstances, PPRL algorithms such as Bloom filters can be susceptible to frequency attacks [88, 98, 99]. Where stronger security guarantees are required, new and/or hardened PPRL methods are needed. Niedermeyer et al. [98] suggest modifications to the construction of basic Bloom filters to improve security, including randomly selected hash values, modification of identifiers, salting, fake injections and multi-field Bloom filters [100]. Schnell et al. [101, 102] have developed a multi-field Bloom variant called Cryptographic Long-term Key (CLK) using different hashing schemes for each field, and some research has also been done into

Secure Multi-Party Computation (SMC) using homomorphic encryption for comparisons [88, 103, 104]. This allows some computations to be performed on encrypted values, producing results that, when decrypted, provides the answer. While homomorphic encryption is seen by some as the Holy Grail for secure computation [105, 106], its use in practice is limited due to very high computation overheads and immature implementations [88, 103, 107] Even with this research into different PPRL techniques, there is still limited research in the context of attacks; more research is required before they can be used to secure sensitive personal data [88, 98, 108].

*Efficiency.* Record linkage can be computationally expensive, and operational databases contain an enormous amount of records which pose tremendous challenges around the scalability of record linkage techniques, especially PPRL protocols. [109]. For PPRL to be operationally feasible, the run times for large scale linkages must be similar to those taken using non-PPRL techniques (i.e. using full demographic information). Of the various privacy preserving algorithms that have been proposed by researchers, the Bloom filter has been shown to be one of the best performers [110].

Distributed computing methods have been employed to solve computing efficiency issues in related areas e.g. entity resolution, covering techniques and Map Reduce based algorithms [111-113]. However, research on distributed PPRL is limited [114, 115]. Karapiperis and Verykios [114] recently presented a framework for PPRL using the Map/Reduce paradigm; however, their evaluation used relatively small datasets (in comparison to real-world data) at 300,000 records and only four local compute nodes. Additional research is required to determine the efficiency of PPRL methods when applied to large operational datasets.

Opportunities exist to trial PPRL methods within scalable, distributed paradigms (i.e. cloud computing). The emergence and uptake of an “Infrastructure as a Service” (IaaS) service model allows for the provisioning of processing, storage and other computing resources, as and when required [116]. The Australian government has actively promoted cloud computing for government, non-profit organisations and research groups, requiring agencies to consider cloud services for new ICT procurements [33]. However, while there is a general “push” towards cloud infrastructure, record linkage centres are not utilising the potential capabilities of this scalable infrastructure. This is due, in part, to the uncertainties and perceived risks associated with locating identifiable data in a cloud environment. Potential exists for PPRL systems to be designed for this infrastructure and utilise the rapid elasticity for on-demand resource usage.

*Accuracy:* Linkage quality metrics assess the ability of a linkage technique to classify records into matches and non-matches correctly [109]. In real world scenarios, high linkage quality is hard to accomplish due to recording errors, missing values, or outdated information in data files. As with traditional linkage methods, a major challenge in the adoption of privacy preserving methods in operational settings is to achieve and maintain high accuracy of results.

Recently, some privacy preserving methods have been shown to achieve high linkage quality [96]; however, the setting was experimental and several issues remained unresolved (i.e. threshold setting, linkage strategy optimisation and quality assessment). These challenges will need to be overcome or adequately addressed before such techniques can be applied to broader operational contexts.

### **3.6. Application of Privacy Preserving Record Linkage**

This research investigated methods for record linkage which do not require full disclosure of demographic information. These privacy preserving methods may be useful for data custodians who are not comfortable or are unable (for legal reasons) to release identifiable information for linkage purposes.

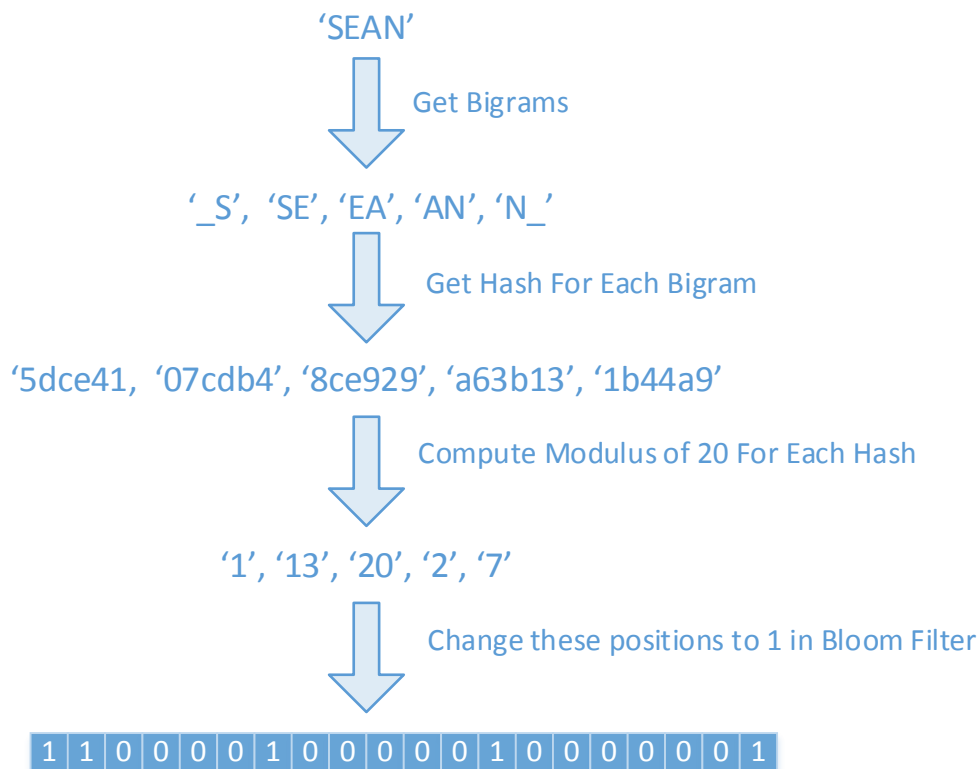
A methodological review of PPRL techniques was carried out, and Bloom filter Privacy Preserving Record Linkage was identified as a compatible option for the PHRN and explored further [99, 117]. The research assessed how this PPRL technique performed with large real administrative datasets. Linkage quality and speed were evaluated in relation to linkage results using fully identified data.

Privacy preserving record linkage using Bloom filters works by encoding personally identifying information into Bloom filters (binary vectors). A Bloom filter begins as an array of a set length, with all elements set to zero. Each field (e.g. first name) is broken down into overlapping sets of letters (qgrams). Padding is often used to give the first and last letters their own bigrams. Each of these qgrams is passed through a series of cryptographic hash functions. A hash function is an algorithm which produces a fixed-length output with several important properties. Firstly, given the same input, it will always produce the same output (i.e. the same qgram will produce the same hash value). Secondly, the hash function is one-way, meaning it is not possible to determine the encoded qgram from any given hash value (i.e. it is irreversible). Different hash passwords can be used to produce different output. In practice, a different password would be used by data custodians for different research

projects. This would provide another layer of security (the same project password would be used by all data custodians' involved in a project to enable linkage to occur).

The modulus of these hashes is then computed with respect to the length of the Bloom filter. This process allows us to map each bigram to a position in the Bloom filter. These positions are then set to 1 (see Figure 8).

**Figure 8: Creating a Bloom filter: a simple example using bigrams**



Two Bloom filters can be compared to each other using a dice coefficient. The Sørensen–Dice coefficient is calculated as twice the number of positions in which both Bloom filters have a value of one, divided by the number of positions set to 1 in total. The dice coefficient results in a score between 0 and 1, where a higher score reflects greater similarity.

$$\text{Dice Coefficient}_{A,B} = \frac{2h}{a + b}$$

where  $h$  is the number of positions set to 1 in both bloom filters,  
 $a$  is the number of bit positions set to 1 in bloom filter A,  
and  $b$  is the number of bit positions set to 1 in bloom filter B.

An example....

Bloom Filter 1: 5 positions set to 1

1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Bloom Filter 2: 6 positions set to 1

1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

$$= \frac{2 \times 4}{5 + 6} = 0.727...$$

The encryption techniques used in privacy preserving linkage with Bloom filters means that probabilistic type techniques can be used during the matching process. These techniques allow for small errors such as spelling mistakes which greatly improve linkage quality. Evaluations using real data found linkage quality using Bloom filters to be equivalent to those achieved using unencrypted personal identifiers, and greater than that of other implemented privacy preserving methods [96].

### 3.7. Conclusion

Record linkage centres nationally and internationally have implemented many different aspects and features to ensure the protection of individual privacy as part of their processes. These include implementation of internationally accepted IT security standards and development of a robust governance framework as well as operational policies and procedures to control data flows and minimise data disclosure risk.

This chapter addresses the fourth aim in the thesis looking at privacy and security around the whole record linkage process. As the demand for record linkage services grows, it is critical for linkage units to implement methods which protect privacy and safeguard national security, yet maximise the benefits that can be derived from administrative and survey data. The book chapter on ‘privacy preserving record techniques in operational linkage centres’ identifies linkage models in existing centres and highlights governance and technical solutions that can be used to minimise disclosure risks.

Improved PPRL methods not only strengthen security but increase linked research opportunities as additional datasets are made available. While many technical challenges of record linkage have been overcome, significant legal and administrative challenges remain which, in turn, impact on data availability and access.

The development of probabilistic linkage techniques that do not require the release of personal information but protect privacy through other mechanisms (e.g. encryption methods) represent a significant breakthrough for data linkage in Australia and internationally. The paper on 'Privacy Preserving record linkage on large real world datasets' describes Curtin's development and evaluation of this technology. There has been significant interest in Curtin's PPRL technology from USA, Canada and the UK.





### 3.8. Book Chapter

**Boyd JH, Randall SM, Ferrante AM. *Application of privacy preserving techniques in operational record linkage centres.* Medical Data Privacy Handbook. Springer International Publishing (2015)**



### 3.9. Published Manuscript

Randall SM, Ferrante AM, **Boyd JH**, Bauer JK, Semmens JB. ***Privacy-preserving record linkage on large real world datasets.*** Journal of Biomedical Informatics (2014)



## Chapter 4

---

### Methods to assess linkage quality - how do we measure up?

*“Politicians use statistics in the same way that a drunk uses lamp-posts - for support rather than illumination”*

*Andrew Lang*

#### *Published Manuscript(s):*

**Boyd JH**, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, Dickinson T. *A simple sampling method for estimating the accuracy of large scale record linkage projects*. *Methods of information in medicine*. 2016;55(3):276-83.

Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. *Use of graph theory measures to identify errors in record linkage*. *Computer Methods and Programs in Biomedicine*. Volume 115, Issue 2, July 2014, Pages 55-63

Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. *The effect of data cleaning on record linkage quality*. *BMC Medical Informatics and Decision Making* 2013; 13 (64): e1-e10.

**Boyd JH**, Ferrante AM, Irvine K, Smith M, Moore E, Brown AP, Randall SM. *Understanding the origins of record linkage errors and how they impact on research outcomes*, ANZJPH; doi: 10.1111/1753-6405.12597.

#### *International Conference presentation(s):*

Ferrante AM, **Boyd JH**, Randall SM, Brown AP, Semmens JB. *How do you measure up? Methods to assess linkage quality*. Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.

Randall SM, Ferrante AM, Brown AP, **Boyd JH**, Semmens JB. *Assessing the impact of different grouping methods: time to rethink and regroup?* Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.



## 4.1. Origins of record linkage error

The linkage process involves creating and comparing pairs of records in order to make a determination about whether they belong to the same person or not. The challenge in designing linkage algorithms is to optimise linkage for data collections being combined. Factors like completeness, consistency, constancy and timeliness of identifying data need to be taken into account when determining the linkage strategy and can affect the accuracy of the final linked datasets.

Linkage error occurs when pairs of records are not complete or include wrong matches, potentially leading to an incorrect grouping of records. There are two types of errors which can occur. False positives (FP) occur when record pairs are incorrectly designated as belonging to the same individual. False negatives (FN) occur when record pairs are incorrectly identified as belonging to different people. The aim in linkage is to maximise the number of true positives (TP) and true negatives (TN).

Using administrative collections and record linkage techniques provides researchers with access to population-based data from across government. These large datasets provide the power to identify relationships with statistical significance. A large number of study participants can be a factor in research design and need to be carefully considered to ensure linked administrative data is analysed appropriately [118]. In addition, it is important that researchers understand the processes around both data collection and linkage to ensure that they are aware of strengths and limitations of data and methods used to bring together records. In this way, the potential for misinterpretation is reduced [119, 120]. This chapter outlines the challenges associated with assessing and reporting linkage quality, the potential impact on the analysis of linked administrative data, and identifies the information required to inform research better.

## 4.2. Linkage quality

In assessing linkage quality, of primary interest is knowing the number of true matches and non matches identified as links and non-links. Any misclassification of matches within these groups introduces linkage errors [121].

The number of incorrect links can be affected by many factors within linkage design, including matching rules or weights and the acceptance threshold. However, the quality of identifiers, in terms of recording errors and completeness, and distinguishing power of identifiers will also have an impact on the matching results.

Achieving high linkage quality is important for ensuring the quality and accuracy of research outputs as well as service and policy reviews based on linked data. Achieving high linkage quality can be difficult and typically requires a large amount of effort. Most linkage systems can be tuned to optimise the false positive and false negative rates. However, all research projects are different, some require a very high degree of accuracy for certain matches, while others require a lesser degree.

#### 4.2.1. Metrics for measuring errors in linkage

Errors in record linkage are usually reported using pairwise precision and recall [26, 122]. These metrics return a number between 0 and 1, where a higher number indicates greater accuracy or 'linkage quality'. Precision refers to the proportion of returned links that are true matches (sometimes called positive predictive value):

$$\text{Precision} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false positives}}$$

Recall is the proportion of all true matches that have been correctly linked. Recall is also known as sensitivity and is measured as:

$$\text{Recall} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

Commonly the harmonic mean of these two values is taken, called the F-measure. By representing the quality of a linkage with just one number, we can more easily compare linkages. The use of the harmonic mean results in higher scores only when both precision and recall have higher scores, unlike a simple average.

$$\text{Fmeasure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

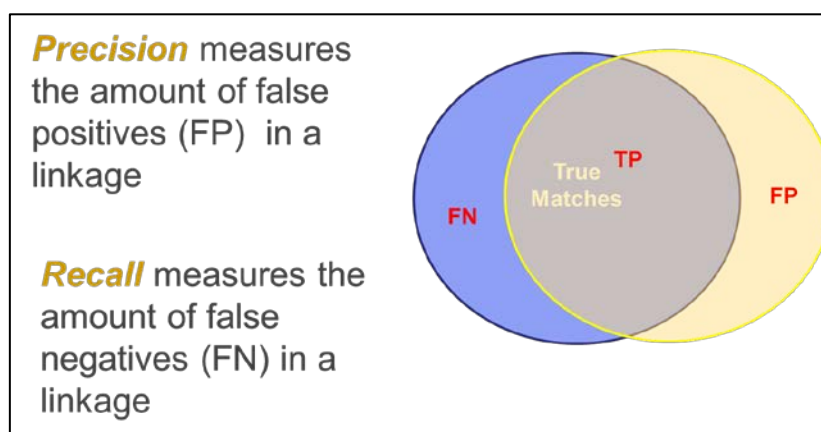
While these three quality metrics are widely used, it is unclear whether these measures are computed consistently. For instance, the recall metric can include as a false negative pair, two records which did not match together but which nonetheless ended up in the same group (i.e. through a third record; an indirect match). As this false negative pair in no way impacted on the final grouping, it is unclear whether it is appropriate to count this as an error.



An alternative method known as saturated Precision, Recall and F-Measure, removes this issue by first grouping the found pairs, and then recreating pairs based on these groups, fully saturating each group with pairs. This method has been used where pairwise information was not available [123]. Any discussion of appropriate metrics should give consideration to the impact of linkage accuracy on research outcomes. Research into the most appropriate metrics for measuring linkage quality will ultimately hinge on our understanding of the relationship between linkage quality and research outcomes.

Group based metrics are harder to define and are often included within linkage quality reports that provide detailed information on records and groups. Linkage quality reports can contain information on group size and consistency as well as expected record combinations i.e. birth registrations with a matching hospital delivery record or in-hospital deaths with a matching death registration.

**Figure 9: Privacy Preserving Constraints**



#### 4.2.2. Estimating linkage quality

Most linkage systems can be tuned to optimise false positive and false negative rates. However, all research projects are different; some require links that are highly accurate while others emphasise maximising linkage rates [121, 124]. Knowing that linkage error can impact on the interpretation of research findings and introduce bias to research studies highlights a need for routinely measuring linkage quality [125]. Although standard methods are available to assess the level of false positive matches produced through linkage, accurately estimating the number of missed matches is not easy [109].

With an ever increasing number of research studies involving linked data, researchers are requesting information on matching quality to ensure the appropriate analyses can be

performed [125, 126]. Developing standard methods to measure matching quality is an important area which can be used by linkage units to refine linkage strategies and inform subsequent analysis; this is essential as the number and complexity of record linkage infrastructure projects continue to expand.

While it is possible to identify false positives based on the results of a linkage (e.g. using targeted clerical review on linkage output), identifying missed links is more challenging and often left unknown [127, 128]. Common quality assurance reporting implementations which contain estimations of false positives and false negatives usually involve complete review of linkage results, 'gold standard' datasets (used as a benchmark to assess linkage quality) and the application of group based logic mapping (e.g. a group of records belonging to a single person which includes a hospital record with a discharge dead code should also contain the associated death registration). However, these techniques are often constrained by the effort involved or the accuracy of the results.

Using a sampling process, it is possible to accurately and consistently estimate linkage quality metrics for large scale (and enduring) record linkage projects. The sampling methodology provides a number of advantages in assessing linkage quality. It offers a manageable and cost effective framework for the assessment of linkage quality (and, additionally threshold setting). By applying this technique, it is possible to assess both the accuracy of matches made as part of a linkage and to estimate the proportion of missed links [128]. The assessment of missed links is traditionally difficult to undertake but can be valuable to researchers who wish to adjust research results based on overall linkage quality.

This method can also be applied to 'deterministic' record linkage, where instead of the probabilistic approach, a series of logical rules are used to determine which records belong to an individual. In the rules-based approach, rules would need to be ordered based on how 'strict' they were – i.e. the likelihood of containing a false positive. Additional rules would also have to be developed, of a lower quality than those currently used, in order to estimate missed matches.

By developing and applying scalable methods of quality assessment, linkage units can assess the accuracy of the matching process and provide research extracts with the appropriate level of linkage quality [125, 129].

### **4.3. Techniques to optimise quality and reduce errors**

The ultimate aim in designing a linkage strategy is to find all possible matches by separating true positives and false negatives into two distinct groups/distributions [11]. However, this is seldom possible, and these groups/distributions often overlap introducing error into the final match selection. As a result, the match selection acceptance process involves balancing the number of false positives and false negatives to maximise overall linkage quality. The challenge is in optimising linkage quality and in knowing how strict to be with the matching criteria. This is important as exacting rules will identify true positives (increasing precision) at the expense of missed matches, leading to a decrease in recall. Linkage units often seek a middle-ground between precision and recall [15].

Linkage units use various techniques to maximise linkage quality. One approach is to optimise the linkage strategy through review and assessment and then undertake targeted interventions of the linkage results. Linkage units will often monitor linkage outputs to confirm consistency of results and to ensure that nothing significant has changed in the data.

#### **4.3.1. Quality and quantity of incoming data**

The quantity and quality of incoming data is known to have a significant effect on linkage quality.

Two studies have shown that linkage accuracy is “strongly dependent on the amount of personal identifying information available on the records being linked”. Newcombe et al (1983) conducted an epidemiological follow-up study of 16,000 uranium mine and refinery employees using a generalised probabilistic record linkage system for searching a national death file [130]. The accuracy of the computerised matching was compared with that of corresponding manual searches of one-eighth of the worker file. The computer was more successful than manual searchers and was also less likely to report false linkages. Newcombe noted that “in both approaches, accuracy was strongly dependent on the amount of personal identifying information available on the records being linked.”

More recently, Bass and Garfield (2002) investigated whether the analysis of data linked by deterministic matching of statistical linkage keys (SLK's) leads to significantly different conclusions than would be obtained through analysis of data linked by probabilistic linkage systems on full demographic data [131]. Two different SLK's were investigated, namely the HACC and SAAP SLK's, and the full demographic data comprised full names, address, date of birth, sex, country of birth and indigenous status [132]. The study showed that “the HACC

and SAAP keys both produce inaccurate linkages compared to that resulting from probabilistic linkage using full demographic data.” Further, statistically significant differences in a number of analyses were found, implying that different linkage methods can lead to significantly varied (and unexpected) results. The authors concluded that “ideally, linkage should be performed using probabilistic methods using as much demographic data as possible.”

Data completeness (extent of missing values) in each variable also has an effect of linkage quality. Most linkage units assess data completeness upon receipt of data from custodians, as part of the standard electronic transfer and load (ETL) processes. Monitoring changes in data completeness over time provides an additional quality check on the status of data collections.

#### **4.3.2. Data cleaning (and standardisation)**

Data cleaning incorporates a variety of different methods which will be appropriate in specific circumstances. The effect of data cleaning depends heavily on the underlying dataset being used, and many techniques may be useful on particular datasets. A key finding from the project was that although data cleaning prior to linkage is common place, it can in some instances be ‘over done’ leading to a decrease in linkage accuracy, due to the reduced variability in each (cleaned) data variable.

Common techniques include:

- *Reformatting values.* Data values can be simply changed to a new format without actually creating or removing information. This can be necessary to ensure all data is in a common standard for comparison during linkage. For example, if two datasets store dates in a different format, these may need to be changed into a common format for comparison. The data are not altered by this transformation, only the representation of the data.
- *Removing punctuation.* Unusual characters and punctuation are removed from alphabetic data items. Names with spaces, hyphens or apostrophes are more likely to be spelt differently, and removing these values will remove this difference.
- *Removing alternative missing values and uninformative values.* Datasets can often contain special coding 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 - for instance hospitals records may contain ‘Baby of Rachael’ in

a forename field, or 'NO FIXED ADDRESS' in an address field. These are commonly removed.

- *Phonetic encoding.* By encoding phonetic information captured in an alphabetic data item (for example surname/family name), names that are given different spellings but sound the same will be brought together. Phonetic encoding is a very common technique that has been used for a long time in record linkage, typically in blocking. Common encoding algorithms include Soundex, NYSIIS and metaphone.
- *Name and address standardisation.* Name standardisation and name parsing are processes used to break down a person's full name into its individual components. For instance, a name field with the entry 'Dr John Harry Williams' can be broken down into its component – title, first name, middle name and last name, with each component being individually compared.

Similarly, an address can be broken down into its basic components such as street number, street name and street type. By creating multiple variables through the standardisation process, small differences between records such as a different order or a slight change 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 large set of rules, but statistical methods have also proved useful.

Less common methods of data cleaning and standardisation include:

- *Sex Imputation.* A record for an individual with a missing sex value can have this value imputed based on their first name e.g. Anna->female; James->male. This requires a lookup table which equates common first names with sex.
- *Development and use of nickname lookups.* A nickname file, containing common nicknames and shortened versions of given names can be used to translate forenames to a common value. Using a nickname lookup, a person recorded as Jim on one dataset and James on another could be given the same first name, potentially bringing these records together.
- *Asian name parsing/segmentation.* Naming conventions in many Asian cultures differ from those used in Western cultures and may be arranged in a different order. Care is required to correctly identify the family name and to correctly distinguish between true given names and other naming conventions. With Vietnamese names, for example, a middle name may indicate:

- a person's gender (Thi as a middle name indicates a female, while Van indicates male);
- a generational distinction (e.g. brothers and sisters may share the same middle name to distinguish them from an earlier generation); or
- a person's position in the family (birth order).

#### **4.3.3. Clerical assessment and review**

Methods of administrative intervention can be used to both evaluate the performance of a linkage strategy and to improve matching quality for specific subgroups [16]. The clerical processes required within a linkage system usually involve methods of validation to ascertain the effectiveness of the linkage design (Clerical Assessment) and manual review of potential matches to confirm links (Clerical Review) [129]. These processes typically involve human assessment of links to assess matching algorithms or to confirm or reject links where the algorithms cannot decide. This manual clerical review process (similar to a chart review) can be based on record pairs or groups of records formed through the linkage. The evaluation of the pairs provides an assessment of many steps in the linkage process including blocking and matching.

#### **4.3.4. Identifying errors to improve linkage quality**

In general, linkage strategies use a standard methodology to all entities within a dataset. Applying a single linkage strategy can disguise issues of heterogeneity within the population. Targeted intervention can be used to improve the overall quality of the links generated by matching algorithms; these can be applied at the 'pair' or 'group' level. Pairs are often targeted based on the strength of the relationship and can be further assessed, manually or automatically, to identify errors and refine links. Alternatively, new and/or existing links can be evaluated by examining all records in a linked group. The final linkage step creates groups of records thought to belong to the same person. These grouping strategies integrate collections of record pairs found through the matching process, to determine the full set of records belonging to the same individual. The characteristics of the group often provide additional information to help identify errors [17]. The chronological review of all records attributed to an individual often tells a story about legitimate changes to identifiers like family name (e.g. through marriage) and addresses, and can be useful in highlighting incorrect links. Automated quality assurance techniques involving the application of "suspicious groups", semantic rules and group theory are all in use by operational linkage units.

#### **4.3.5. One-off assessments of linkage quality**

Although it is possible to identify false positives based on the results of a linkage (e.g. using targeted clerical review on linkage output), identifying the missed links is more challenging and often left unknown [18]. One solution involves access to a 'gold standard' dataset that can be used as a benchmark to assess linkage quality. These 'gold standard' datasets can be either real or synthetic but must allow identification of all 'correct matches' to become the truth set against which a linkage strategy can compare results.

#### **4.3.6. Sensitivity analysis**

Matching results can be affected by parameters, comparators and thresholds defined within the linkage model (or linkage design). Sensitivity analysis allows linkage units and researchers to evaluate potential changes to linkage design (including acceptance thresholds) and whether these changes have an impact on research conclusions (depending on the analytical model applied). This process allows identification of any potential bias and allows adjustment for linkage error in the final statistical analysis ensuring valid research findings.

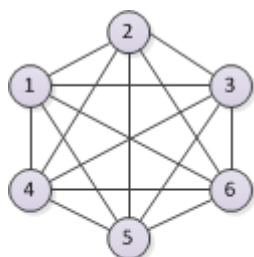
#### **4.3.7. Graph theory – Picturing the problem**

Graph theory is the examination, identification and understanding of mathematical structures which model pairwise relations [133]. The fundamental structure of interest is a "graph" which is made up of a collection of 'nodes' and lines called 'edges' which join one node to another. These pair relationships exist in record linkage, and are the building blocks used to determine groupings of records that are considered to belong to the same person. In the person-based record linkage context, nodes are individual records, with edges representing the record-pair associations found through the linkage process.

The similarities between record linkage and graph theory have been noted previously (Huang, 2006). Work has been carried out in translating probabilistic record linkage practices into graph theoretic language (Lenz, 2003). The purpose of this research was to evaluate whether constructs from graph theory could be usefully applied to identify incorrect groups of records arising from record linkage.

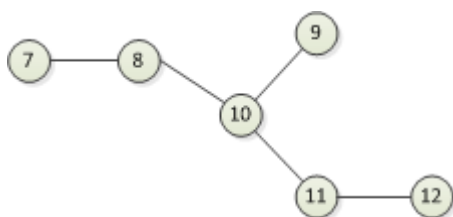
The investigation looked at three graph theory metrics (completeness, diameter and bridges) which were successfully adapted and used to identify groups of records containing errors. These graph-theory methods were able to target incorrect groups very accurately (i.e. high PPV); however, they could only identify a small percentage of all incorrect links (low sensitivity).

**Figure 10: Graph theory metrics**



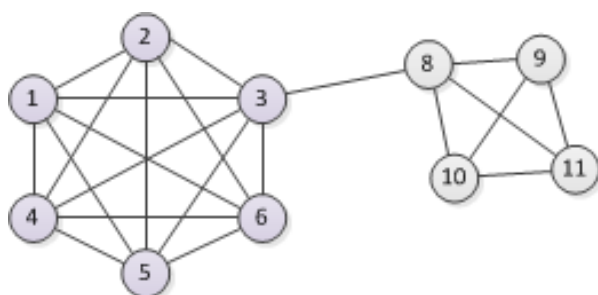
**Completeness**

Fully connected or “saturated” groups = all records match to all others



**Diameter**

Longest distance from one record to another



**Bridges**

An edge (match) which, if removed, causes a disconnection

In practical terms, these graph theory techniques show promise as an additional method of identifying suspicious or incorrect groups in record linkage output. Other graph theory techniques have already been included in matching applications like ChoiceMaker, which has a transitivity engine to enhance grouping.

**4.3.8. Automated quality tools**

As linkages and linked files increase in size the amount of clerical checking even within selected groups increases. A number of programs have been developed to improve linkage quality through automated or semi-automated processes. These ‘quality tools’ select records of interest (based on user-defined criteria or rules), check to see if they contain problems to fix or resolve linkage issues. If the linkage issues cannot be resolved, the records are added to a clerical file, for manual review by an operator.

The software takes groups of records that have been brought together in the matching criteria and evaluates them against a set of internal consistency rules. Based on the outcome of this evaluation, the groups are dynamically adjusted. The tool can be used to supplement the clerical review process, and reduce clerical load to a manageable amount.



The tool has been designed to be fast, scalable and requires minimal intervention by operators.

These programs have been developed as external components of the linkage system. Decisions from these quality review processes are input into the system via the 'batch quality review' envelopes.

#### **4.3.9. Rule-based clerical intervention**

Using expert information gained from the clerical review process, a series of rule-based checking algorithms have been developed to reduce manual clerical burden on large scale linkage systems [134]. Evaluations of these algorithms have been included to identify whether there is value developing these further.

The methodology is designed to replicate the type of rules used by reviewers to determine whether records belong to the same person. The automated assessment applies logical rules to decide if the record pairs are 'links' or 'non links'. These logical rules are held outside the system and can be modified, removed or added to by operators [135].

The rules-based assessment methodology uses an iterative process allowing clerical reviewers to identify additional rules which may be added to the logic to supplement the already available rules and further automate the clerical review process.

The automated assessment is 'trained' for each linkage by one (or more) of the reviewers. The reviewer's knowledge is added and validated incrementally based on their manual clerical review of pairs.

#### **4.4. Impact of linkage quality on research outcomes**

While several papers have observed the affect linkage quality can have on research results [26, 27], little is known about how linkage errors (and different types of errors) directly impact on specific methods of analysis [136, 137].

There is some evidence that an increase in false positive errors in a one-to-one linkage of registry and death information is likely to lead to an underestimate of survival, while false negative errors lead to overestimation [28]. One-to-one linkage (involving only two datasets, where each has one record per person) is a relatively simple design in which matching can often be clerically assessed; the relationship between errors in more complex linkage scenarios is often difficult to evaluate. 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 bigger role in biasing outcomes is a major factor which would help the interpretation and validity of research findings.

There is also some evidence that record linkage errors are not distributed evenly throughout the population. Instead, these vary among particular subgroups. Subpopulations with greater levels of linkage error include women [29, 30], the elderly [30], ethnic minorities [30, 31], defined geographic areas (from recording differences in particular localities) and those from lower socioeconomic groups [32]. Analysis of both linked and unlinked records is an important step that allows researchers to assess potential variations within population subgroups e.g. geographical, cultural, remoteness, etc.

#### **4.5. Mitigating the impact of linkage error**

Linkage errors are extremely tough for researchers to detect, as they typically have no access to personally identifying information, and must rely on the accuracy of linkage keys provided by data linkage units. However, it is important researchers realise that linkage errors may introduce bias in studies. The risks include systematic error that may arise from the linkage design, quality assurance or from threshold setting.

Usually, researchers become aware of errors as a result of early application of consistency or 'sanity' checks, where combinations of content information attributed to one individual are found to be invalid. Consider, for example, a hospital record of a full term delivery occurring after a hospital record indicating a hysterectomy. Either there is an error in recording this clinical information, or these two records do not belong to the same individual.

Methods to control for linkage error within linked analysis are currently being developed. Chambers and colleagues have developed a series of methods for including estimated linkage errors provided by linkage units as additional factors within regression analysis [138-140]. However, these have so far focused on linkage errors arising from simple one-to-one linkage. Several other methods for including linkage error within regression analysis utilising Bayesian statistics have been proposed [141, 142]. An approach, led by Goldstein utilises all record-pair associations, along with their confidence, in one-to-one matching (rather than only the highest) [143]. This approach introduces linkage error variation into the estimation of clinical variables, which are then used in statistical analysis. However, methods to allow researchers to control for linkage error in the more typical many-to-many linkage scenarios are yet to be developed.

Given the researcher's limited ability to detect incorrect links and the infancy of statistical methods to control for linkage error, it is vital that linkage units work with researchers to develop sound statistical models and to provide accurate and detailed information about the quality of the links provided. Information on linkage quality allows researchers to assess/address any bias in the study design (e.g. if data is coming from different systems are the data and linkage results consistent [144, 145]) or to allow adjustment to statistical confidence levels in the interpretation of results.

#### **4.6. Conclusion**

Record linkage is a powerful technique which allows this data to be transformed from discrete episodes relating to specific service contacts at distinct points in time into complex pathways providing information on an individual's interactions with services over extended periods of time. Achieving high linkage quality is essential for ensuring and maintaining the quality and integrity of research and related outputs based on linked data.

It is important that researchers make time to understand both the data being used within a study (i.e. how it was collected, the coding structure, how complete etc.) and the linkage process used to create the participant profiles for a record linkage study. This may require additional information from data linkage units such as reports on the software, linkage strategy and on matching quality to ensure the appropriate analyses can be performed [25, 39].

As opportunities for international and cross sectorial data linkage studies increase, there is a need to understand differences between data collections, sub-populations and linkage results. Greater transparency and improved reporting of linkage results can be used to help researchers improve study design, understand the impact of analytical techniques and strengthen the interpretation of results. At the moment there are no standard metrics for assessing and reporting on the quality of linkage outputs.

This chapter provides a summary of new and innovative technologies, developed as part of this research, which can be used to measure and improved linkage quality (fifth aim).



#### 4.7. Published Manuscript(s)

**Boyd JH**, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, Dickinson T. ***A simple sampling method for estimating the accuracy of large scale record linkage projects.*** Methods of information in medicine (2016)



Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. *Use of graph theory measures to identify errors in record linkage.* Computer Methods and Programs in Biomedicine (2014)





Randall SM, **Boyd JH**, Ferrante AM, Semmens JB. ***The effect of data cleaning on record linkage quality.*** BMC Medical Informatics and Decision Making (2013)



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 '11<sup>th</sup> 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,

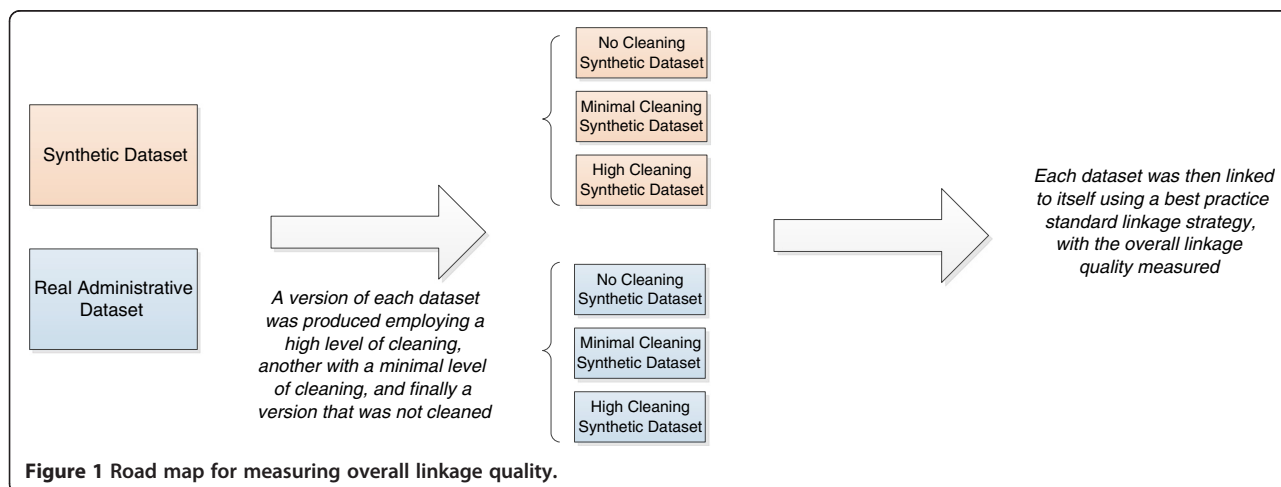
the relative improvement of each transformation on the above datasets was measured and averaged Figure 1.

**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



**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
<b>Reformat values:</b> Not required	<b>Reformat values:</b> Not required	<b>Reformat values:</b> Not required
	<b>Remove alt. missing values and uninformative values:</b> Invalid dates of birth removed Invalid postal code values removed	<b>Remove alt. missing values and uninformative values:</b> Invalid dates of birth removed Invalid post code values removed
	<b>Remove punctuation:</b> Both forename and surname fields had all punctuation and spaces removed	<b>Remove punctuation:</b> Both forename and surname fields had all punctuation and spaces removed
		<b>Nickname lookup:</b> Nicknames were changed to their more common variant.
		<b>Sex Imputation</b> 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
<b>Reformat values:</b> Date of birth reformatted.	<b>Reformat values:</b> Date of birth reformatted	<b>Reformat values:</b> Date of birth reformatted.
	<b>Remove alt. missing values and uninformative values:</b> Invalid dates of birth were removed Invalid postcode values were removed ('9999' etc.) Uninformative address and suburb values removed ('NO FIXED ADDRESS', 'UNKNOWN' etc.) Birth information encoded in first name removed ('TWIN ONE OF MARTHA' etc.)	<b>Remove alt. missing values and uninformative values:</b> Invalid dates of birth were removed Invalid postcode values were removed ('9999' etc.) Uninformative address and suburb values removed ('NO FIXED ADDRESS', 'UNKNOWN' etc.) Birth information encoded in first name removed ('TWIN ONE OF MARTHA' etc.)
	<b>Remove punctuation:</b> Forename, middle name surname and suburb fields had all punctuation and spaces removed	<b>Remove punctuation:</b> Forename, middle name surname and suburb fields had all punctuation and spaces removed
		<b>Nickname lookup:</b> 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 measured 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.

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

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

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

**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	- <sup>a</sup> 0.08%	+0.08%
Remove alt. missing values	+0.5%	0%
Nickname lookup	-28%	-33%
Sex Imputation	NA	-5%

<sup>a</sup> 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 increase 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
<i>Percentage difference from original variable</i>			
Given name original	0.006575	0.946085	0.013059
Given name with removed punctuation	0.006573↓ <sup>b</sup> 0.03%	0.947188↑ <sup>b</sup> 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%

<sup>b</sup> 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.naacr.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 edition US Patent 1261167; 1918.
22. Taft RL: **Name search techniques.** New York: Bureau of Systems Development; 1970.
23. Phillips 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: SIGKDD; 2003.
29. Christen P, Churches T, Hegland M: **Febri—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.

#### 4.8. Published Letter

**Boyd JH**, Ferrante AM, Irvine K, Smith M, Moore E, Brown AP, Randall SM. *Understanding the origins of record linkage errors and how they impact on research outcomes.* ANZJPH (2016)



## Chapter 5

---

### National Data Linkage - Proof of Concept

*“However beautiful the strategy, you should occasionally look at the results”*

*Winston Churchill*

#### **Published Manuscript(s):**

**Boyd JH**, Randall SM, Ferrante AM, Bauer JK, McInnery K, Brown AP, Spilsbury K, Gillies M and Semmens JB. Accuracy and completeness of patient pathways - the benefits of national data linkage in Australia. *BMC health services research* 15.1 (2015): 312.

Spilsbury K, Rosman D, Alan J, **Boyd JH**, Ferrante A, Semmens JB. *Cross border hospital use: An analysis using data linkage across four Australian states*. *Medical Journal of Australia* 202.11 (2015): 582-586.





## **5.1. Proof of Concept – Research using linked data**

The Proof of Concept (PoC) collaboration projects were designed to demonstrate the capability of using PHRN infrastructure to address research questions of national significance by undertaking inter-state linkages. The PoC collaboration was initially envisaged as a single study and was eventually designed as a collection of four carefully selected and complementary research studies intended to assess different components of the PHRN network [146].

Each of the PoC collaboration projects was extensive and complex involving datasets from various jurisdictions. As a result, the approvals process required agreement from multiple data custodians and human ethics committees as well as data transfer agreements executed with each data custodian. Each project had specific technical, logistical and epidemiological objectives.

## **5.2. Indicators of hospital mortality**

The initial PoC demonstration project was the first in Australia to link person-based inpatient and mortality data across jurisdictions (i.e. across PHRN nodes located in different states) [147]. The project served the dual purpose of developing and testing the data acquisition and linkage processes needed to support ongoing multi-state research of this kind, and to contribute valuable information to a public health issue of national importance - mortality during or within 30 days of an inpatient hospitalisation. The epidemiological aims of this project were consistent with the Australian Commission on Safety and Quality in Health Care (ACSQHC) program of work to develop national indicators for hospital mortality.

## **5.3. Project phases**

There were two key stages to the PoC#1 project. The first involved the development of national linkage keys and involved negotiating governance arrangements and access to data before the creation of the linkage map from provided data:

- The transfer of hospital and mortality personal identifiers from NSW, WA, SA and QLD to the CDL;
- The linking of this data to create a national map; and
- The transfer of this map back to jurisdictions.

The second stage involved analyses of linked data to fulfil the epidemiological aims of the research project. This component of the project sought to answer research questions of national importance through the construction and analysis of inter-state linkages.

## 5.4. Data linkage

Up to ten years of hospital discharge and death registration data from four Australian States were included in the PoC collaboration - including data from WA, NSW, SA and QLD. Hospital morbidity data was provided from both public and private hospitals in WA, NSW and QLD; only public hospital admission data were available for linkage from SA hospitals at the time of the study.

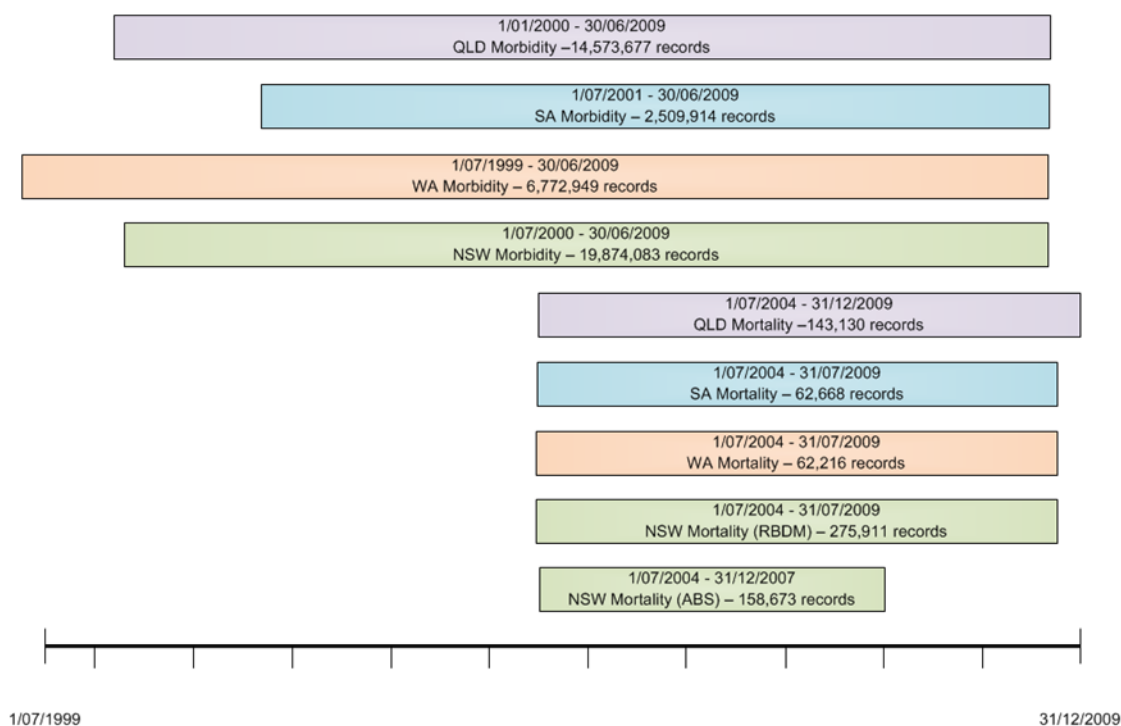
### 5.4.1. Data summary

Over 44 million records were included in the linkage process for this PoC Collaboration (Table 1 and Figure 11). Applying the separation model within the project, only personally identifying demographic information was provided for linkage. The information available for the matching process included full name (first name, middle name and family name), gender, date of birth and address, along with admission and separation dates for hospital events (or date of death, for mortality events).

**Table 1: Summary of datasets and number of records**

<b>Data Collection</b>					
	<b>WA</b>	<b>NSW</b>	<b>SA</b>	<b>QLD</b>	<b>Total</b>
<b>Hospital</b>	6,772,949	19,874,083	2,509,914	14,573,677	43,730,623
<b>Death</b>	62,216	275,911	62,668	143,130	702,598
		158,673			
<b>TOTAL</b>	<b>6,835,165</b>	<b>20,308,667</b>	<b>2,572,582</b>	<b>14,716,807</b>	<b>44,433,221</b>

**Figure 11: Datasets provided to CDL**



During this proof of concept project, the CDL created person-based linkage keys across the PHRN nodes using common demographic variables. The CDL standardised the data linkage technique used across nodes, and as a result produced different linkage results compared with each jurisdiction performing their linkage independently. The project enabled the CDL to refine its operational models based on the availability and quality of demographic data and linkage keys provided by the collaborating PHRN nodes [144].

#### **5.4.2. Linkage strategy**

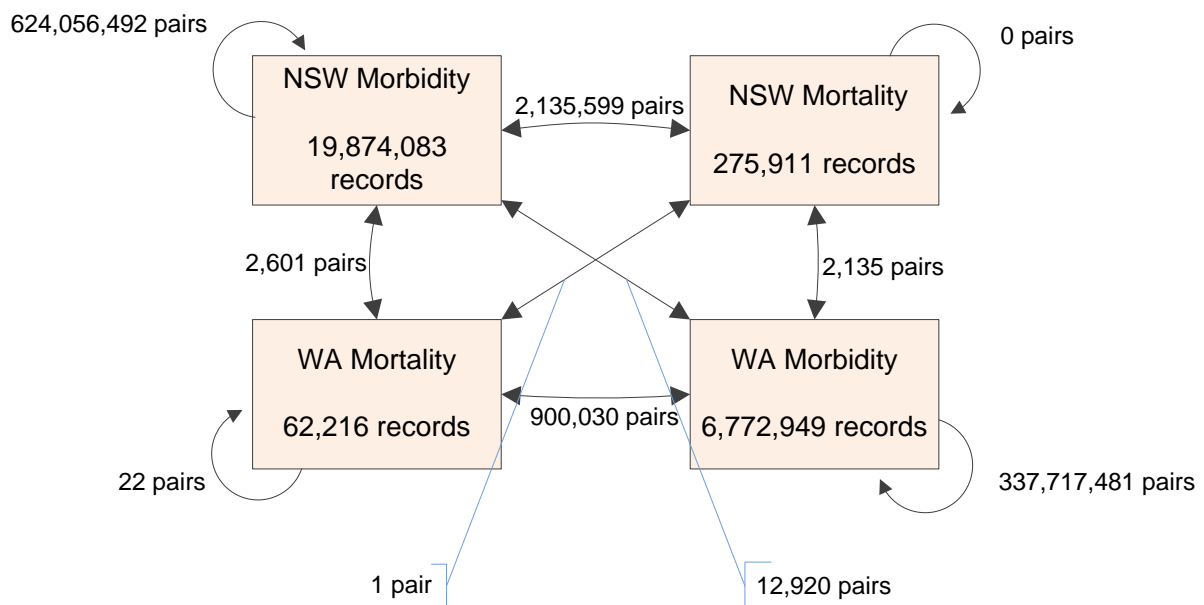
The linkage strategy employed followed a typical probabilistic record linkage approach [148]. This matching process consisted of a sequence of comparisons between two records followed by a judgment about whether the two records belong to the same individual [16, 22, 23]. The linkage strategy included a 'blocking' step which limited comparisons to those records who shared a minimum level of identifying demographic information.

#### **5.4.3. Cross-jurisdictional data linkage results**

Prior to this project, the CDL linked both NSW and WA data as new and compared their results to those achieved by the WA Data Linkage Branch (WADLB) and the NSW Centre for Health Record Linkage (CHeReL). This evaluation linkage was undertaken to confirm the accuracy of processes and to refine the linkage strategy.

The evaluation linkage process involved internal linkage of records within datasets (de-duplication) as well as linkages across multiple datasets. The output from the linkage process consisted of record pairs. Figure 12 illustrates the number of pairs found in the evaluation linkages of WA and NSW data alone. A total of 964,770,789 matched pairs were identified.

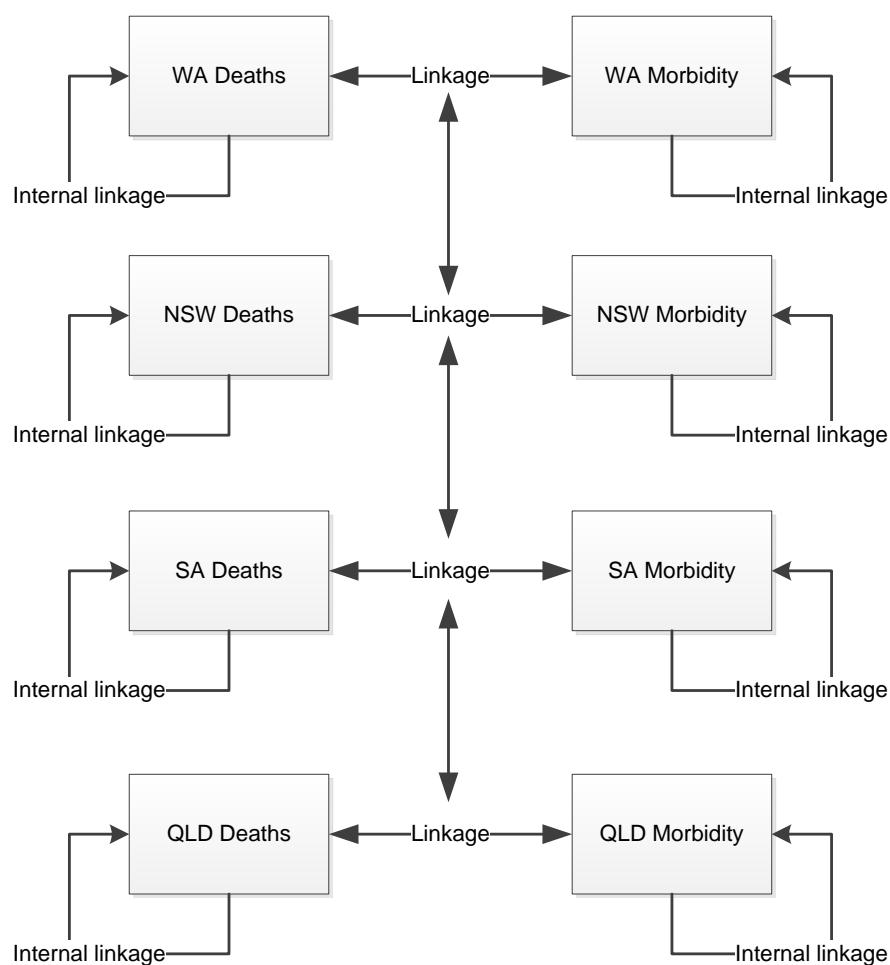
**Figure 12: Number of pairs found in each linkage**



The evaluated matching strategy was then applied to all cross-jurisdictional datasets (WA, NSW, SA and QLD). The matching strategies varied according to the nature of the datasets being brought together and the characteristics of records within the datasets.

The linkage process involved both the internal linkage of records within each dataset (de-duplication) and linkage across multiple datasets (see Figure 13). The linkages were performed using purpose built linkage software (NLS) and infrastructure established by the CDL.

**Figure 13: Nature of data linkage of WA, NSW, SA and QLD data**



The final results of linkage across the various jurisdictions are summarised in Table 2. From the 44 million records supplied for linkage, the matching process identified 12 million discrete entities. Forty five per cent (45%) of these individuals were identified as having had a single hospital admissions record; the rest had an average of 5.9 hospital records per person.

The proportion of individuals with a single hospital record was different in each of the four jurisdictions: Western Australia had the smallest percentage of people with a singleton hospital record (35%) and South Australia with the highest (52%). In the same way, the hospital records per individual (i.e. average group size) ranged between 6.2 and 5.2 in WA and SA respectively. The differences in the South Australian figures could be influenced by the absence of private hospital records from the state.

**Table 2: Patient summary results**

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

Notes:

\* Individuals who only have one hospital record in their group.

\*\* Singletons are not included in the total number of individuals for this calculation

Hospital admission statistics and cross-border population flow over the study period are summarised in Table 3. Individuals with hospital records from a single state only were classified as a 'static population', individuals with records in more than one state were classified as a 'mobile' population. The proportion of individuals classified as 'mobile' was largest in QLD with 5% of individuals having hospital records in other states and smallest in SA and WA (with 3% of individuals having hospital records in other states). This 'mobile' population accounted for between 4% and 7% of the hospital records in each state.

**Table 3: Patient mobility**

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

**Notes:**

<sup>^</sup>Mobile population refers to the number of individuals in a jurisdiction/state that have records in other states.

<sup>^^</sup>Static population refers to the number of individuals in a jurisdiction/state that have records ONLY in that state.

#### **5.4.4. Blocking efficiency**

As part of the linkage strategy, a sequence of blocks was applied which aimed to lower the number of linkage comparisons without affecting the linkage quality (i.e. reduce comparisons without missing 'True Positive' links). The reduction ratio and pairs completeness score were calculated to evaluate the efficiency and effectiveness of the block.

The reduction ratio assesses the drop in pair comparisons resulting from a blocking strategy. It measures the efficiency of a blocking strategy without evaluating the impact on linkage quality. The reduction ratio is calculated as the ratio of blocked comparisons resulting from a blocking strategy to the total possible comparisons.

The proportion of 'true pairs' blocked or pairs completeness metric can be used to assess the impact of blocking on linkage quality. The pairs completeness metric measures the number of 'true positive' pairs compared in the blocking strategy as a percentage of all possible true positive pairs. In this study, the true pairs were identified using the jurisdictional linkage keys for WA and NSW records.

There is a natural trade-off between the reduction ratio and percentage of 'true pairs' blocked. If we reduce the comparisons for efficiency it has an impact on linkage quality; similarly, maximising quality by increasing the number of comparisons can have a considerable impact on the time required to process the linkage. As a result, the blocking strategy is a reference point for all linkage quality estimates (i.e. precision and recall). If a comparison is excluded by the blocking strategy it will never create a link.

Using the blocking strategy outlined, approximately 142 billion comparisons were performed during the linkage process. These comparisons made up only 0.014% of all possible record pairs in the whole comparison space. Within each jurisdiction the blocking process was similar, with the state based reduction ratio ranging between 0.99973 and 0.99987. Table 4 provides a summary of comparisons generated.

**Table 4: Blocking efficiency**

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

\* 'True' pairs based on the jurisdictional linkage key supplied by WA and NSW. Jurisdictional linkage keys from SA and QLD were not available at time of project.

#### 5.4.5. Comparison of CDL Linkages with Jurisdictional Linkages

Along with hospital and mortality records from SA, NSW and WA, the CDL were provided with jurisdictional linkage keys. These linkage keys identified records which belong to the same person, as determined by that state linkage unit. Access to this information enabled the CDL to compare between the links created at a national level to those identified by the state unit. These jurisdictional linkage keys allowed us to gauge the ability of the national system to link very large datasets to a high quality in a short period of time.



These comparisons rely on the use of the jurisdictional linkages as the 'gold standard'. While it is possible that incorrect and missed links exist in jurisdictional links, it is assumed that such errors are minimal, as the data has been linked to a very high standard. These links have been developed and checked over a long period of time, with extensive manual clerical reviews performed. These links have also been validated by researchers who have used them widely. Significant expertise has been developed by these organisations which have a long history of linkage.

#### **5.4.6. Linkage accuracy**

Jurisdictional links from WA and NSW were used as a 'gold standard' allowing an evaluation of the PoC project linkage quality against each individual state (that is, comparing within-state results against jurisdiction links provided by WA and NSW).

Results from the PoC linkage were compared against links produced by state-based linkage units in WA and NSW using the jurisdictional linkage key supplied with each state dataset. The accuracy of the linkage was exceptionally high with a very small number of pairs identified by WA and NSW jurisdictional linkage keys being lost as a result of the blocking strategy (Table 5). Using the assigned linkage strategy, over 99.67% of all 'true pairs' were made available for comparison through blocking. The number of blocked 'true pairs' provided a baseline for assessing the linkage quality.

Using the WA data less than 0.1% of hospital pairs identified as links were found to be incorrect and over 98.1% of all possible within-jurisdiction links were found. This trade-off between precision and sensitivity resulted in a maximum F-measure quality score of 0.99 (where 1.000 would indicate a perfect linkage) indicating 'an average' error rate for morbidity data from these jurisdictions of less than 1%.

Missing data was a major factor which had an effect on both blocking and matching accuracy. Just under one-third (30%) of NSW hospital records did not contain any name information (these records were supplied from private hospitals). The impact was substantial with the overall quality of the linkages using WA data better than that of NSW. The linkage of hospital morbidity records in NSW provided an overall F-measure of 0.976 (precision = 98.8% and recall = 96.3%).

The NSW results were also broken down by hospital status (public versus private). Linkage results for the public hospitals showed better concordance with existing CHeReLs links (F-Measure = 0.995) than the private hospital data. The private data has less identifying

demographic information for linkage which relates to a drop in linkage quality (F-Measure = 0.949). See Table 5.

Linkage quality was highest for data from Western Australia. These results are reflective of good data quality and established high-quality linkage processes in the various states.<sup>1</sup>

**Table 5: Summary of linkage quality (results for NSW and WA)**

	NSW			WA
	Morbidity	Public	Private	Morbidity
<b>Precision</b>	0.988	0.994	0.983	0.999
<b>Recall</b>	0.963	0.996	0.917	0.981
<b>F-measure<sup>#</sup></b>	0.976	0.995	0.949	0.990

**Notes:**

# F-measure is the harmonic mean of precision and recall.

In general, the results show the CDL were cautious in their approach, accepting fewer incorrect links at the cost of more missed links.

The results for linkages on WA data were higher than that of NSW. New South Wales results were further examined by measuring the linkage quality for only those records with name information (Public Hospital records) compared to the linkage quality of only those records without (Private Hospital records). Records with name information showed much higher results (F-Measure = 0.995) indicating the lack of named private hospital data is the most likely reason for the drop in quality.

Along with the above pair based quality metrics, groups were also checked to investigate the spread of errors. For Western Australian groups, 98.7% contained no incorrect records, while 97.6% of groups contained all legitimate records for that group. For New South Wales groups, 98.6% contained no incorrect records, and 84% contained all of the legitimate records for that group.

<sup>1</sup> In NSW and Queensland, name information from public hospitals is not released for data linkage. This places a limit on what can be achieved with data linkage. Insufficient or poor quality data will degrade linkage quality.

## 5.5. Epidemiology

Using linked data, the research team investigated whether person-based morbidity data produced a more refined estimate of risk-adjusted hospital standardised mortality ratios (HSMRs, both in-hospital and within 30 days of discharge) compared with HSMRs calculated using separation-based morbidity data.

An important goal in improving the quality of care provided by hospitals is to avoid unnecessary deaths. In Australia, many of the standard reports around hospital-related deaths focus on post-surgical events. However, there is an increasing need nationally and internationally to be able to look at deaths due to any cause to get a picture of a hospital's performance overall. This assessment of quality used together with other clinical performance indicators provides a good platform to access and enhance hospital performance, accountability and probity.

The PHRN Proof of Concept Collaboration #1 is unique in Australia being one of the first research projects to link hospital separation data with hospital-related deaths data across different states. The primary epidemiological aim focussed on deaths that occurred in hospital or within 30 days of hospitalisation in Western Australia, Queensland, New South Wales and South Australia.

## 5.6. Conclusion

The PoC comparison project showed that national linkage of 'big data' can be carried out efficiently and accurately using the infrastructure developed by the CDL. The study clearly demonstrated the importance and impact of cross-jurisdictional linkage. It allowed researchers to understand population movements better and to assess health service utilisation across State borders at an individual or person-based level by linking various disparate datasets from different government organisations. The impact of more complete patient pathways on research outcomes has not been previously documented and is not well understood. This project provided reliable estimates of cross-border population flows and service utilisation for the first time.

The 'blocking' strategy applied in this linkage was shown to be effective, with a large number of comparisons being removed from the matching process with very little impact on linkage quality. Using existing validated linkage keys from WA and NSW to evaluate the project linkage, little difference was found between links created for the project and those found by the jurisdictional linkage units in both WA and NSW.

This PHRN PoC collaboration project demonstrated the feasibility of efficient and accurate large scale data linkage (Aim 6). The project produced high quality data linkage results without the need to apply comprehensive manual quality review procedures, which would have been resource intensive and extremely costly on a dataset of this size. The most significant outcome was the ability of this linkage to identify cross-border population movement, providing researchers with information to fully describe patient pathways.

## 5.7. Published Manuscript(s)

**Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInnery K, Brown AP, Spilsbury K, Gillies M and Semmens JB. *Accuracy and completeness of patient pathways - the benefits of national data linkage in Australia.* BMC health services research (2015)**



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 McInnery, Adrian P. Brown, Katrina Spillsbury, 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.

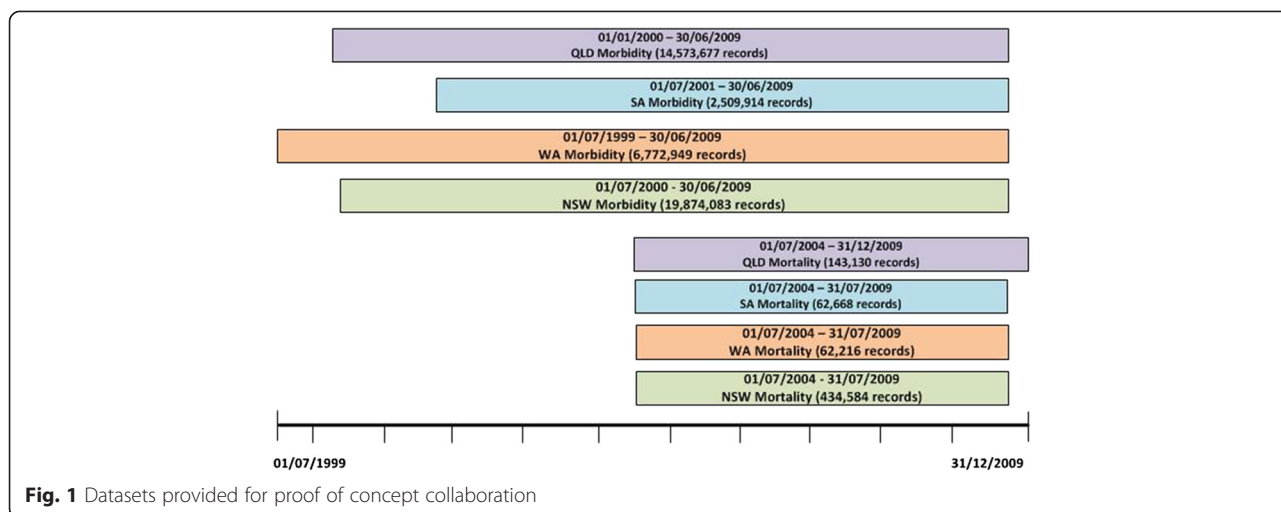
The second aim of the paper is to demonstrate the importance and impact of cross-jurisdictional linkage. The study will capture population movement at individual or person-based level through linkage of disparate datasets, enabling researchers to assess the full extent of health service utilisation across 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





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 (%) <sup>a</sup>	99.76 %	99.95 %	-	-	-

<sup>a</sup>'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).

### 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 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 <sup>a</sup>	0.976	0.995	0.949	0.990

<sup>a</sup>F-measure is the harmonic mean of precision and recall

**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 <sup>a</sup>	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 <sup>b</sup>	5.4	6.2	5.2	5.9	5.9

<sup>a</sup>Individuals who only have one hospital record in their group

<sup>b</sup>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 <sup>a</sup>	205,551	47,575	29,645	202,859
% of individuals in that state	4 %	3 %	3 %	5 %
Static population <sup>b</sup>	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 %

<sup>a</sup>Mobile population refers to the number of individuals in a jurisdiction/state that have records in other states

<sup>b</sup>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, Kliever 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](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.

Spilsbury K, Rosman D, Alan J, **Boyd JH**, Ferrante AM, Semmens JB. **Cross border hospital use: An analysis using data linkage across four Australian states.** Medical Journal of Australia (2015)





## Chapter 6

---

### Using National Record Linkage Infrastructure to Support Research

*“Research is creating new knowledge”*

*Neil Armstrong*

#### **Published Manuscript(s):**

**Boyd JH**, Wood FM, Randall SM, Fear MW, Rea S, Duke JM. *Effects of pediatric burns on gastrointestinal diseases: A population-based study*. The Journal of Burn Care & Research, (2016) doi: 10.1097/BCR.0000000000000415.

Duke JM, Bauer J, Fear MW, Rea S, Wood FM, **Boyd J**. (2014). *Burn injury, gender and cancer risk: population-based cohort study using data from Scotland and Western Australia*. BMJ open, 4(1), e003845.4

#### **International Conference presentation(s):**

**Boyd JH**. *Using record linkage to examine long-term effects of burn injury: The Western Australian Population-based Burn Injury Project*. Conference: International Population Data Linkage Conference, Swansea, Wales, August 2016.



## **6.1. Data availability and use**

Collection, storage and management of data have come a long way in the last two decades with significant developments in technology regarding power and capacity [149]. The volume of digital data is increasing exponentially, providing more available data for research [6]. This increase in information also includes routinely collected administrative data which provides the building blocks for critical analysis used to shape policy, evaluate performance, allocate resources and improve public services [150].

Private and public organisations collect significant amounts of data pertaining to their business processes and services. Analysing these individual and aggregate records provides an opportunity to improve knowledge of the environment. Data manipulation and analytics can unlock the potential within the data and combining data from different sources can often provide a better understanding of the 'big picture'.

Using a whole system approach recognises that many interacting factors can influence individual parts of the system and that solutions to problems have to be developed taking these factors and interactions into account. Data linkage is a way of bringing together data to provide information on the whole population, generating a more complete picture of the community than is possible using other research methods [45, 151-153]. It is also a very cost-effective research tool.

Research using linked routine administrative data has demonstrated its value, with access to health, education and criminal justice datasets crucial in supporting policy and improving public services [154]. To demonstrate the benefits of using linked data in research I have included two case studies. These important clinical research programmes highlight the translation of innovative record linkage infrastructure into research outcomes.

## **6.2. Research translation case studies**

The power of linkage systems and the value of building routinely linked data repositories comes from the research outputs this infrastructure enables [8, 51]. Two case studies are presented from Scotland and Western Australia to highlight the importance of linked data to the research community. Both these case studies provided research evidence from linked data to support and change clinical practice ensuring better patient experiences and outcomes.

In Scotland, availability of a linked repository of morbidity and mortality data allowed calculation of national clinical outcomes. These were used by clinical audit groups in the Scottish Government as part of an assessment framework for Scottish hospitals [155]. Clinical indicators are firmly established as a core output within the NHS in Scotland and have helped generate a wider programme of work which includes patient safety, benchmarking and surgical mortality [153, 156].

The clinical outcome indicators around heart disease also generated interest from a research group made up of clinicians, public health professionals and researchers. This group used the linked data in Scotland to explore many aspects of heart disease, changing clinical practice, patient pathways and to open up additional research areas. The research papers from these studies are still widely cited by heart disease researchers.

The final case study focuses on burn injury research in Western Australia. This research project used data from both Western Australia and Scotland to explore the risk of cancer following burn injury. The study used data from Scotland to confirm results obtained from an initial study using linked data from Western Australia [157]. Using data from Scotland improved statistical power and allowed researchers to study site specific cancers following burn injury. This study was part of a larger programme of work exploring the impact of burn injury and the risk of subsequent morbidities following a hospitalised burn. This information is particularly important in informing the primary care treatment of patients following a burn injury.

### **6.3. Research using linked data in Scotland**

Publication of clinical outcomes is now common practice in many health systems around the world. These have their roots in work carried out in the United States where public reporting of performance indicators was routinely published as far back as the early 1990's [158-160]. However, Scotland led the way in Europe around the production and public release of clinical indicators.

In 1992, the Clinical Outcomes Working Group was set up as a Sub Committee of the Clinical Resource and Audit Group, within The Scottish Office. The group was tasked with producing comparative clinical outcome indicators for Scotland using linked administrative data. National clinical indicators were first produced in Scotland in 1993. Since then, Scotland has been at the forefront of producing national indicators and in exploring potential for more detailed research on specific morbidities [156].

*S Capewell, S Kendrick, J Boyd, G Cohen, E Juszczak, J Clarke. Measuring outcomes: one month survival after acute myocardial infarction in Scotland. Heart 1996; 76(1):70-5. DOI: 10.1016/S1062-1458(97)82162-9*

One of the first clinical indicator research studies looked at one month survival after acute myocardial infarction to see if it could be a useful means of measuring outcome of hospital care. The research explored acute myocardial infarction as an outcome indicator and examined survival effects after adjusting for available prognostic factors such as age, sex, co-morbidity, deprivation and deaths outside hospital.

The study identified 40,371 admissions to hospital with a principal diagnosis of acute myocardial infarction (ICD9 code 410) during 1988-1991. The research project looked at both in hospital survival at 30 days (77%) and overall survival at 30 days (53%) when 18,452 acute myocardial infarction deaths in the community were included. Applying logistic regression to model survival in the study cohort, we found that the odds of dying within 30 days:

- almost doubled for each decade of age;
- were (slightly) higher in females than in males;
- nearly doubled in patients with a previous history of infarction, coronary heart disease, or other heart disease; and
- were significantly increased in patients with circulatory disease, respiratory disease, neoplasm, or diabetes.

From this study, one month survival after acute myocardial infarction was identified as a useful measure of hospital care for patients admitted with an AMI. Even after adjusting for prognostic factors, marked variations in survival between different hospitals and health areas persisted. It was noted that these differences could be accounted for by other factors and the study generated further research. In addition, it was agreed that (where possible) in future studies AMI survival outcomes should take account of infarct severity.

Research Translation: Access to high quality linked data has provided a platform to develop a range of clinical indicators which are routinely produced as part of national statistics outputs in Scotland. These reports are used by the National Health Service to monitor performance, allocate resources and evaluate changes to clinical practice.

### **6.3.1. Studies Linking ISD Data for Epidemiology (SLiDE)**

Scottish linked data has also been used to extend the understanding of disease groups, exploring patient treatment and pathways to understand the burden in the population better. One such project, Studies using Linked ISD Data for Epidemiology (SLiDE), used the linked datasets to examine trends, prognosis and deprivation effects in Coronary Heart Disease (CHD) patients (initially funded by the British Heart Foundation). A number of later studies focused on coronary artery bypass grafting surgery, atrial fibrillation, unstable angina, chest pain and the burden of CHD in primary care.

***K MacIntyre, S Capewell, S Stewart, JWT Chalmers, J Boyd, A Finlayson, A Redpath, JP Pell and JJV McMurray. Evidence of improving prognosis in heart failure: trends in case fatality in 66 547 patients hospitalized between 1986 and 1995. Circulation 2000;102:1126-1131 doi: 10.1161/01.CIR.102.10.1126***

The paper addresses the challenges of generalising results from clinical trials which had shown a reduction in case fatality related to heart failure with therapies such as angiotensin-converting enzyme (ACE) inhibitors and beta-blockers in middle-aged men recruited to the study.

This retrospective cohort study used the linked health administrative data from the Scottish morbidity collections and death registrations. The cohort was selected as heart failure patients admitted to a Scottish hospital between 1986 and 1995 (using the International Classifications of Diseases, 9th Revision 425.4, 425.5, 425.9, 428.0, 428.1 and 428.9). The study used a patient's first heart failure record in the period as the 'index' admission. Patients with a hospitalisation related to heart failure in the previous five years were excluded.

Hospital and death data from the Scottish morbidity collection and death register were linked to the heart failure cohort in the study period (1986 to 1995). Baseline data for each patient included information on age, gender, date of admission, date of death (if it occurred), geographic location (postcode sector) and deprivation indices (Carstairs Deprivation category (1 to 5) from the 1991 census data). As well as crude case-fatality rates (univariate analysis); Kaplan Meier plots (median survival), Cox proportional hazard regression and logistic regression (at 30 days, one year, five years and ten years) were used to explore survival.

The adjusted case-fatality rates for the heart failure cohort confirmed the effect of age on long term case fatality (from 30 days to end of follow-up period). With the hazards ratio per decade of age 1.42 men and 1.38 for women, the effect of sex in the model was modest with highly significant interaction between age and sex. Deprivation increased both short-term (26% in men and 11% in women) and long-term (10% in men and 6% in women) case fatality.

Over the study period (1986-1995) median survival increased from 1.23 to 1.64 years. After adjustment for age, sex, deprivation and prior admission, the short-term (30 days) case-fatality rates for heart failure patients fell by 26% in men (95% confidence interval (CI): 15-35) and by 17% in women (95% CI: 6 to 26). Longer-term case fatality rates fell by approximately 18% in men (95% CI: 13 to 24) and 15% in women (95% CI: 10 to 20).

The study identified demographic differences in heart failure patients selected from the linked dataset (i.e. using the whole heart failure community) to those enrolled in clinical trials, with the community cohort containing more elderly patients and a great proportion of females. The data from the study also showed that the prognosis for patients admitted to hospital was worse than indicated by clinical trials.

The study confirmed a very poor prognosis for patients admitted to hospital for the first time with a diagnosis of heart failure. However, the study showed that case fatality in patients admitted to hospital with heart failure was falling over the period with plenty room for further improvement.

***K MacIntyre, S Stewart, S Capewell, JWT Chalmers, J Boyd, A Finlayson, A Redpath, H Gilmour, JJV McMurray. Gender and survival: a population-based study of 201,114 men and women following a first acute myocardial infarction. Journal of the American College of Cardiology Volume 38, Issue 3, September 2001; Pages 729-735 doi: 10.1016/S0735-1097(01)01465-6***

With the independent effect of sex on the prognosis of patients with Acute Myocardial Infarction (AMI) uncertain, the study was designed to address residual issues concerning gender based differences in AMI. Using linked data, the study looked at gender based differences in immediate case fatality of AMI patients and whether any differences persist in the long term. The study hypothesised that a higher proportion of men dying before reaching hospital would help explain women's excess short term fatality.

The study cohort was selected as AMI patients admitted to a Scottish hospital between 1986 and 1995 (using the International Classifications of Diseases, 9th Revision 410). The study used a patient's first AMI record (hospital or death) in the period as the 'index' record. Patients with a hospitalisation related to AMI in the previous five years were excluded.

Hospital and death data from the Scottish morbidity collection (SMR01) and death register were linked to the AMI cohort in the study period (1986 to 1995). Baseline data for each patient included information on age, gender, date of admission, date of death (if it occurred), geographic location (postcode sector) and deprivation indices (Carstairs Deprivation category (1 to 5) from the 1991 census data). Chi-squared tests for categorical and t-tests for continuous data were performed. Survival analysis was conducted using Kaplan Meier method and multiple logistic regression at 30 days.

Of the 201,114 individuals with a known first AMI between 1986 and 1995 in Scotland, on average, women were seven years older than men. 117,749 of these individuals (58.5%) survived to be admitted to hospital (41.5% were AMI death registrations with no associated hospital admission).

The adjusted case fatality rates following hospital admission for first AMI showed that the short-term case fatality in young women (under 55 years of age) was higher than age-matched men (Odds Ratio (OR) 1.25 (1.07-1.46)). With increasing age this disparity in 30 day survival was reduced with men and women over 65 years of age having equivalent case fatality.

In contrast to the higher case fatality rates seen among younger women *following* admission, men were more likely to die *before* hospitalisation. AMI death before reaching hospital in young women (under 55 years of age) was lower than age-matched men (OR 0.86 (0.79-0.94)). Excess deaths occurring out of hospital in men offset the excess risk of death following admission to hospital in younger women (OR 0.94 (0.87-1.01)).

In the adjusted longer term survival analysis, no difference was seen between men and women in case fatality in those individuals admitted to hospital (OR 0.97 (0.93-1.01)).

At the time, this was the largest population-based study to examine age and gender based differences at short and long term mortality following admission with a first AMI. With conflicting evidence on the effect of gender on short term case fatality in previous research, the results showed for inpatients admitted to hospital, the odds of death were 12% greater



for women than men for every 10 year decrease in age. However, when deaths from AMI that occur outside hospital are taken into consideration, the 30-day case fatality rates are greater in men than in women. Thus, accounting for varying case fatality rates before hospitalisation seems to partly explain the gender-based differences observed in short-term case fatality in patients admitted to the hospital.

***S Capewell, K MacIntyre, S Stewart, JWT Chalmers, J Boyd, A Finlayson, A Redpath, J Pell, JJV McMurray. Age, sex, and social trends in out-of-hospital cardiac deaths in Scotland 1986-95: a retrospective cohort study. The Lancet 11/2001; 358(9289):1213-7. DOI: 10.1016/S0140-6736(01)06343-7***

Acute Myocardial Infarction (AMI) accounts for roughly three quarters of all coronary heart deaths and half of all hospital admissions for coronary heart disease. Half of these deaths are sudden deaths occurring outside of hospital and without any admission. Few studies have been able to examine such trends because most studies are confined to hospital admissions. This study looked at trends in death rates out of hospital, stratified by sex, age, and socioeconomic status, to identify inequalities and enable prioritisation of future interventions.

This retrospective cohort study used linked health administrative data from the Scottish morbidity collections and death registrations in the study period (1986 to 1995). We identified all deaths in the 117,718 patients with acute myocardial infarction who were admitted to hospital, and all 83,365 people who died out of hospital i.e. individuals who did not survive to reach hospital after a first acute myocardial infarction (using the International Classifications of Diseases, 9th Revision 410).

Hospital and death data from the Scottish morbidity collection (SMR01) and death register provided information on age, sex, date of admission, date of death, geographic location (postcode sector) and deprivation indices (Carstairs Deprivation category (1 to 5) from the 1991 census data). These data were used to derive population-based mortality rates, stratified by sex, age, socioeconomic status, and year. Out-of-hospital mortality rates were calculated for men and women of different age groups and deprivation categories. Multivariate logistic regression was performed to examine the effect of year of first acute myocardial infarction on probability of reaching hospital.

Out of hospital deaths accounted for 69.2% (95% CI 69.0–69.5) of all AMI deaths. Between 1986 and 1995, 44,655 men and 38,710 women had a first acute myocardial infarction and

did not survive to reach hospital. Overall, the number of deaths out of hospital per year fell substantially from 9,484 to 6,712.

Generally, population-based rates of out-of-hospital mortality increased with age. Out-of-hospital deaths as a proportion of all acute myocardial infarction events, increased with age from 20.1% (19.2-21.0) in patients younger than 55 years, to 62.1% (61.3–62.9) in those older than 85 years.

The age adjusted population-based mortality rates for deaths out-of-hospital fell by a quarter in women and by more than a third in men. Mortality rate falls were much larger in younger age groups, with a 5.6% average yearly fall in men aged 55–64 years, which was more than twice the 2.5% average fall in men older than 85 years (3.8% vs 2.5%, respectively in women). In addition, population-based mortality rates were substantially higher in deprived socioeconomic groups than in affluent groups.

With an overall fall in coronary heart disease mortality, it is easy to miss inequalities in some of the sub-population groups. This large unselected cohort describes coronary mortality trends in an entire population and allowed us to explore inequalities in age, sex and socioeconomic class. The study showed that women, elderly people and deprived groups had been left behind and suggested that these inequalities should be actively addressed by prevention strategies.

### **6.3.2. SLiDE research summary**

As evidenced by various publications using linked data, Scotland has created robust linkage infrastructure that provides a platform for undertaking routine national linkage while meeting the requirement to provide a secure and controlled environment for working with sensitive data.

Research Translation: The existence of a permanently linked file in Scotland has allowed researchers to exploit administrative data and develop innovative research studies which access the linked file. As a result, linked research capacity has been increased by removing the need to relink datasets for each study.

Accessing and extracting from the linked Scottish morbidity and mortality file allows efficient and cost effective research. The national infrastructure allows data from a diverse and rich range of health datasets to be linked permanently. This resource enables nationally and

internationally significant population level research, to improve health and wellbeing and enhance the effectiveness and efficiency of health services in Scotland.

## **6.4. Burn injury research in Western Australia**

Burns are a distressing injury associated with both physical and mental health conditions for patients over an extended period of time after wound healing. The long-term health and treatment of these patients are related to the powerful metabolic, inflammatory, immune and endocrine changes in the body following burn injury that can last for at least three years.

Western Australia has a long history of burn research which has been used to support and inform patient treatment. Epidemiological evaluation of patient pathways, health care utilisation and clinical outcomes has helped influence the model of care for burn injury in the state. The data resulting from the population-based burn injury project have been used directly to develop clinical guidelines for models of care for burn injury care in Western and to inform the Australia and New Zealand Burn Association, the peak body for health professionals responsible for the care of the burn injured in Australia and New Zealand.

The Burn Injury Research Unit at the University of Western Australia is a national research unit that undertakes basic scientific, clinical and epidemiological research of burn injury to improve the quality of life of the patient. Understanding the link between the basic systemic responses to injury and clinical outcomes is crucial in developing innovative clinical solutions to ensure better patient outcomes.

### **6.4.1. Burn injury and cancer risk in Western Australia and Scotland**

Results of a previous population-based study of burn injury and cancer risk revealed that female burn survivors had an increased risk of all-cause cancer [157]. However, Western Australia has a population of approximately 2.2 million, and as such, did not support detailed age, gender and site-specific cancer incidence assessments with adequate statistical power. Use of parallel datasets from Scotland, of population approximately 5.5 million, allowed examination of the consistency of results and trends across the populations.

The research explored the risk of cancer after burn injury using a retrospective longitudinal study design based on linked hospital morbidity, cancer and death data of persons hospitalised for burn injury in both WA and Scotland. The study aimed to, firstly, confirm the increased risk of 'all-cause' cancer in WA female burn survivors using the Scottish data; and, secondly, examine site-specific cancer risk amongst survivors of burn injury.

Using data from both Western Australia and Scotland, this study identified consistent trends of increased cancer incidence amongst female burn patients for many selected types of cancer. Further experimental and clinical studies are required to enable better understandings about the role of gender in the immune response to burn injury and possible mechanisms of a causal pathway to cancer, which may then elucidate possible sites for intervention. Findings of this study would suggest that it may be prudent for patients with burns to be monitored for development of poor health outcomes, including malignancy, in addition to acute burn follow-up, and that research of the potential therapeutic interventions be investigated.

#### **6.4.2. Western Australian Population-based Burn Injury Project**

The international burn injury and cancer risk study was part of the initial work to investigate long-term health effects caused by burns. A burn research project, the Western Australian Population-based Burn Injury Project was developed and led by Associate Professor Janine Duke in response to clinical questions raised by Professor Fiona Wood, burns surgeon and Director of the Burns Service of Western Australia. This research programme was approved by the health research ethics committees of the Department of Health, Western Australia and the University of Western Australia. The project incorporates a multidisciplinary research team that includes epidemiologists, burn clinicians, health scientists and economists, to investigate increased long-term morbidity and mortality for paediatric and adult burn trauma patients.

Through a retrospective cohort design, the study used Western Australian population-based linked hospital data to explore morbidity in a burn and uninjured population group. All the investigations were performed using a de-identified extraction of hospital morbidity records for all burn patients admitted to Western Australian hospitals with a first burn injury between 1<sup>st</sup> January 1980 and 30<sup>th</sup> June 2012. Burn injury for each patient profile was defined using the International Classification of Disease (ICD) 9 codes 940 to 949 or ICD 10 AM codes T20 to T31. An index event was identified as the first hospital admission in a patient's medical record history in which burn injury, defined by these ICD codes, was recorded as the principal diagnosis or additional diagnosis.

A population-based comparison cohort was randomly selected from Western Australian Birth Registrations (<18 years) and Electoral Roll (≥18 years); any person with an injury hospitalisation during the study period was excluded from the population-based non-injury cohort by WADLS. The resultant comparison cohort (i.e. no injury hospitalisations) was

frequency-matched 4:1 on birth year, gender and year of index burn discharge for the period 1980–2012.

Morbidity and mortality data from the Western Australia Hospital Morbidity Data System (HMDS) and Death Register were linked to the burn and comparison cohort for the study period (1980–2012). Hospital separation data (discharges) included principal and additional diagnoses, age and gender, Aboriginal status, date of admission, dates of discharge or separation and mode of separation. Indices of social disadvantage (Socio-economic Indices for Areas (SEIFA)) and remoteness index (Accessibility Remoteness Index of Australia (ARIA+)) were supplied by Western Australia Data Linkage Branch for both the burn and comparison cohort. SEIFA scores were partitioned into quintiles to generate five ordinal categories from most disadvantaged to least disadvantaged. ARIA+ indices were used to classify geographical disadvantage and access in terms of distance from services by five remoteness categories: major cities, inner regional, outer regional, remote and very remote Australia. The death data included cause of death and date of death. Cause of death was classified using ICD9-CM and ICD10-AM disease and external cause codes.

Burn injury characteristics (total burns surface area percent (TBSA%)) were defined using supplementary diagnostic codes ICD9-CM 948 or ICD10-AM T31. These diagnostic codes were used to classify patients into those with minor burns (<20% TBSA), severe burns ( $\geq$ 20% TBSA) and burns of unspecified TBSA.

Chi squared tests for categorical and Kruskal–Wallis tests for non-parametric continuous variables were performed with the level of significance set at 0.05. Kaplan–Meier plots of survival estimates for burn vs. non-injury and for burn severity (minor, severe, burns unspecified TBSA vs. non-injury) were generated and log rank tests were used to compare the survival distributions of burn and no injury cohorts. The impact of burn injury on long-term survival was estimated using Cox proportional hazard regression adjusting for potential confounders (index age, gender, Aboriginal status, comorbidity, social disadvantage and remoteness, prior hospital use, record of congenital anomaly and index year). The hazard ratios (95% confidence intervals (CI)) estimated from the Cox proportional hazards model were used as measures of Mortality Rate Ratios (MRR). Attributable Risk Percent (AR%) was used to estimate the proportion of long-term mortality where burn injury was an attributable cause. The AR% was calculated as the adjusted rate ratio minus one, divided by the adjusted rate ratio, multiplied by 100 ( $AR \% = ((adjMRR - 1)/(adjMRR)) \times 100$ ). The percentage of deaths in the burn injury cohort that was attributable to burn injury was estimated after adjusting for known potential confounders.

Research of post burn morbidity, using multivariate negative binomial regression and Cox proportional hazard regression analyses, identified increased cardiovascular, musculoskeletal, respiratory and gastrointestinal morbidity in terms of post burn hospital admissions. In addition, evidence generated by 'basic scientific' research at the Burn Injury Research Unit (BIRU) strongly implicates immune changes after both minor and severe burn injury leading to increased susceptibility to infection, cancer, bone loss and cardiac changes. These findings of increased mortality and morbidity relate to both severe and minor burns in the study period. This is a significant finding as the majority of burn injury admissions in Australia, as for other developed countries, are for minor burn injuries.

Health and economic outcomes from this research have been used to affect change in health policy around burn services in Western Australia. Professor Fiona Wood, co-investigator and the Director of the Burn Service of Western Australia (BSWA) has presented research results to the Chief Medical Officer, Executive Director of Public Health and the Injury and Trauma Clinical Network, Department of Health Western Australia. The information has been used to inform strategic decisions about policy, clinical practice and prevention strategies to reduce burn injury and subsequent post burn morbidity and hospital costs.

#### **6.4.3. Effects of paediatric burn injury on gastrointestinal diseases**

One specific research project looked at the effects of paediatric burn injury on gastrointestinal diseases. As with the other Western Australian Population-based Burn Injury Project investigations, administrative data in the study covered the whole of the Western Australian population which allowed long-term analysis of an important and vulnerable study group (i.e. children younger than 15 years when hospitalised for a first burn injury) to a level of detail not permitted by sample surveys. Severity of digestive disease in the study is classified as serious enough for admission to hospital (and we also account for pre-existing and co-morbid conditions).

The analysis showed that the number of post-burn digestive admissions reduced over the study period reflecting improvements in treatment and changes in hospital referral patterns. However, the difference between gastrointestinal admission rates in the burn and comparison cohorts increased over the 33 year period. The burn cohort gastrointestinal admission rates were highest for conditions relating to the oesophagus, stomach and duodenum which had a 5-fold greater admission rate than the comparison group over the study period.

Previous studies have also shown that increased permeability of the intestinal tract following a burn injury affects the intestinal barrier function and this may be a factor in the elevated long-term post burn morbidities. This increase in hospital admission rates (and length of stay) for gastrointestinal conditions in the paediatric burn injury cohort suggest prolonged effects of burn on the digestive system. We recognise that the results may underestimate the impact of burn injury on ambulatory digestive conditions but in the absence of reliable linked primary care data, we believe the findings still have high clinical importance and implications for longer-term patient monitoring.

## **6.5. Conclusion**

In Scotland, the existence of permanently linked national data and facilities for linkage has increased the demand for linked analysis and new linkages. The new linkages have consisted primarily of matching external datasets of various forms - survey data and clinical audit datasets - to the central holdings. Other specialised linkages have involved extending the linkage of subsets of the ISD data holdings back to 1968 for epidemiological purposes (for example, Grampian Lifestyle Survey [161] and MIDSPAN [162]).

In addition, thousands of linked analyses have been carried out ranging from simple patient based counts to complex epidemiological analyses. Among the major projects based on the linked datasets have been clinical outcome indicators (published at hospital level on a national basis), analyses of trends and fluctuations in emergency admissions and the contribution of multiply admitted patients.

In Australia, the burn trauma research collaboration has explored different aspects of burn injury using linked data in Western Australia. This epidemiological research has identified that the long-term health impacts of the burn injury require careful management during both the period of immediate care in specialist burns units as well as subsequent primary/outpatient care following discharge.

These research programs show the value of large responsive linkage infrastructure which can support population-based research. These projects show how linked data can be used to support the translational pathway from population-based research to clinical practice.





## 6.6. Published Manuscript(s)

**Boyd JH**, Wood FM, Randall SM, Fear MW, Rea S & Duke JM. *The Effects of pediatric burns on gastrointestinal diseases: A population-based study. Journal of Burn Care & Research* (2016)



Duke JM, Bauer J, Fear MW, Rea S, Wood FM, **Boyd J. *Burn injury, gender and cancer risk: population-based cohort study using data from Scotland and Western Australia.*** BMJ Open (2014)



# BMJ Open Burn injury, gender and cancer risk: population-based cohort study using data from Scotland and Western Australia

Janine M Duke,<sup>1</sup> Jacqui Bauer,<sup>2</sup> Mark W Fear,<sup>1</sup> Suzanne Rea,<sup>1,3</sup>  
Fiona M Wood,<sup>1,3,4</sup> James Boyd<sup>2</sup>

**To cite:** 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. doi:10.1136/bmjopen-2013-003845

► Prepublication history for this paper is available online. To view these files please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2013-003845>).

Received 19 August 2013  
Accepted 19 December 2013



CrossMark

For numbered affiliations see end of article.

**Correspondence to**  
Professor Janine M Duke;  
[janine.duke@uwa.edu.au](mailto:janine.duke@uwa.edu.au)

## ABSTRACT

**Objective:** To investigate the risk of cancer and potential gender effects in persons hospitalised with burn injury.

**Design:** Population-based retrospective cohort study using record-linkage systems in Scotland and Western Australia.

**Participants:** Records of 37 890 and 23 450 persons admitted with a burn injury in Scotland and Western Australia, respectively, from 1983 to 2008. Deidentified extraction of all linked hospital morbidity records, mortality and cancer records were provided by the Information Service Division Scotland and the Western Australian Data Linkage Service.

**Main outcome measures:** Total and gender-specific number of observed and expected cases of total ('all sites') and site-specific cancers and standardised incidence ratios (SIRs).

**Results:** From 1983 to 2008, for female burn survivors, there was a greater number of observed versus expected notifications of total cancer with 1011 (SIR, 95% CI 1.3, 1.2 to 1.4) and 244 (SIR, 95% CI 1.12, 1.05 to 1.30), respectively, for Scotland and Western Australia. No statistically significant difference in total cancer risk was found for males. Significant excesses in observed cancers among burn survivors (combined gender) in Scotland and Western Australian were found for buccal cavity, liver, larynx and respiratory tract and for cancers of the female genital tract.

**Conclusions:** Results from the Scotland data confirmed the increased risk of total ('all sites') cancer previously observed among female burn survivors in Western Australia. The gender dimorphism observed in this study may be related to the role of gender in the immune response to burn injury. More research is required to understand the underlying mechanism(s) that may link burn injury with an increased risk of some cancers.

## INTRODUCTION

The burden of burn injury is a significant issue, not only in terms of the acute care perspective but also the chronic health effects resulting from the injury event and the subsequent treatments. While burns predominantly

## Strengths and limitations of this study

- Population-level linked administrative data minimise the issues of selection and reporting bias, and loss to follow-up.
- Consistency of results using population-level data from two patient populations provides a greater support for link between burn injury and some cancers.
- Unable to examine the impact of burn severity on cancer risk due to lack of available data.

affect the skin, burns are associated with significant systemic effects,<sup>1–3</sup> depressed immune functioning<sup>4–8</sup> and prolonged periods of systemic catabolism and hypermetabolism.<sup>9</sup> Prompted by a clinical scenario of a diagnosis of hepatocellular carcinoma in a young male burn patient,<sup>10</sup> and the potential for malignancy after burn, an initial study of burn injury and total (all sites) cancer risk was undertaken.<sup>11</sup>

Results of our initial study demonstrated a gender effect with female burn survivors having an increased risk of all types of cancer.<sup>11</sup> In contrast to our results, a Swedish population-based study<sup>12</sup> that linked burn patient hospitalisation records and cancer registrations reported the risk of developing any form of cancer for combined gender was increased (standardised incidence ratio (SIR), 95% CI 1.11, 1.06 to 1.16), with the risk of lung cancer also significantly increased (SIR, 95% CI 1.39, 1.21 to 1.59). However, a Danish study of burn patients<sup>13</sup> found no significant increases in cancer (combined gender) for all malignant neoplasms, including all skin cancers (SIR, 95% CI 0.99, 0.93 to 1.06).

This previous study of burn injury and cancer risk used population-based data linked by the Western Australian Data

Linkage System (WADLS).<sup>14</sup> Western Australia has a population of approximately 2.2 million, and as such, did not support detailed gender-specific and site-specific cancer incidence assessments with adequate statistical power. While the WADLS has linked datasets of Western Australians since the 1970s, other Australian states have only recently established record linkage systems and were not able to support this study. To extend the population base and enable further detailed examination of the observed gender effect and site-specific cancer incidence, approval was sought to access the Scottish population-based record linkage system, the Information Service Division (ISD) Scotland,<sup>15</sup> that has routinely linked health data since the 1980s.

This article reports a retrospective longitudinal study to explore cancer risk after burn injury using linked hospital morbidity, cancer and death data of persons hospitalised for burn injury in Western Australia and Scotland. The study aimed to, first, confirm the increased risk of total ('all sites') cancer observed in the preliminary Western Australian study of female burn survivors using the Scottish data; and, second, examine site-specific cancer risk among survivors of burn injury.

## METHODS

Study data were obtained from the WADLS<sup>14</sup> and the ISD (Scotland) of the National Health Service (NHS) National Services Scotland<sup>15</sup>. The WADLS and ISD Scotland are validated record linkage systems that routinely link administrative health data from core datasets for the entire population of Western Australia and Scotland, respectively.

An index burn injury was defined as the first hospital admission with a burn injury using primary and additional diagnosis International Classification of Diseases (ICD) codes 940–949 (ICD9) and T20-T31 (ICD10). A deidentified extraction of all linked hospital morbidity records (Hospital Morbidity Data System (HMDS), mortality (Death Register, Western Australia) and cancer records (Western Australia Cancer Registry (WACR)) for all persons admitted to hospital with an index burn injury in Western Australia, for the period 1 January 1983 to 31 December 2008, was performed by WADLS.<sup>16</sup> A corresponding deidentified extraction of all linked hospital morbidity records (Scottish Morbidity Records (SMR) 01), cancer registrations (SMR06) and death records (General Register Office for Scotland (GROS)) using the same burn cohort definition was undertaken by ISD Scotland. Hospital admissions data items included age at admission, gender, admission date, separation (or discharge) date, principal and additional diagnoses and external cause of injury.

The WACR was established in 1981 and is a population-based cancer registry based on mandatory notification of cancers from pathologists, haematologists and radiation oncologist, and cancer information from death records.<sup>17</sup> Malignant cancers are coded according

to a modified Australian version of the International Classification of Diseases, Tenth Revision (ICD10-AM) and International Classification of Diseases for Oncology (ICD-O-3).

The Scotland Cancer Registry (SMR06) has recorded all incident cancers in Scotland from 1958, and since 1997 registration has been centralised at ISD Scotland.<sup>18</sup> The registry is responsible for the collection of information on all new cases of primary malignant neoplasms, carcinoma in situ (including grade III intraepithelial neoplasia), neoplasms of uncertain behaviour and (since 1 January 2000) benign brain and spinal cord tumours arising in residents of Scotland. Data quality is monitored using routine indicators, computer validation and ad hoc studies of data accuracy and completeness of ascertainment.<sup>19 20</sup> The Scottish cancer notifications are coded using the ICD V.10 and the ICD-O.

Methods for analysis have been previously published.<sup>11</sup> An incident cancer was defined as a cancer diagnosis notification (C00-C96, excluding C44) after hospital admission for index burn injury. Analysis was restricted to malignant neoplasm notifications (C00-C96, excluding C44) for which total (all sites) and site-specific cancer incident rates were provided by WACR and ISD Scotland for respective populations. Records were excluded from the analysis if the date of cancer diagnosis was prior to date of discharge for index burn hospitalisation. When a record was identified as having more than one malignant neoplasm notification, each neoplasm was counted as an individual record; however, if multiple tumours of the skin (C43) with identical morphological characteristics (ie, the first three digits ICD-O-3 morphology code) were identified, they were recorded only once. Gender and age-specific cancer (total C00-C96, excluding C44) incident rates for the Western Australian and Scottish populations were pooled for the calendar periods 1983–1988, 1989–1993, 1994–1998, 1999–2003 and 2004–2008 to allow for changes in population cancer incidence during the study period.

For the determination of incident rates, the calculation of person-years began on the day of final hospital discharge for the index burn admission, and the study observation period continued until date of the defined cancer diagnosis, death or 31 December 2008, whichever occurred first. Individual calculations were conducted for total (all sites) and site-specific cancers. The observed number of cases of cancer and person-years at risk were calculated by age (5-year age groups), gender and calendar period (1983–1988, 1989–1993, 1994–1998, 1999–2003 and 2004–2008). The expected number of cancer cases was estimated by multiplying the specific number of person-years per category by the corresponding incidence of cancer in Western Australia, Scotland, and combined cancer incidence rates, provided by WACR and ISD Scotland. SIRs were calculated by dividing the observed number of cases by the number expected.<sup>21 22</sup> The 95% CIs were defined under

the assumption that the observed number of cancers followed the Poisson distribution.<sup>23</sup>

Separate SIR analyses for total (all sites) and site-specific cancers were conducted using country-specific data for respective burn patient cohorts (all burn depth) hospitalised from 1983 to 2008; total (all sites) SIRs were repeated for subcohorts of burn admissions from 1983 to 1987, with an optimum follow-up time. To further explore the gender impact of burn injury on cancer risk, total (all sites) cancer SIR analyses were repeated on age-restricted subcohorts classified to reflect the reproductive age at admission for burn injury: <15; 15–49 and ≥50 years. All statistical analyses were performed using Stata statistical software V.11 (StataCorp LP, College Station, Texas, USA).

## RESULTS

As previously reported, in Western Australia from 1983 to 2008, there were 23 450 hospital index admissions for burn-related injury.<sup>16</sup> After exclusion of records with a history of cancer prior to separation date or death during hospital admission for burn, a total of 22 705 patient records were included in the analysis.<sup>11</sup> There were 673 patients with a first cancer notification after the date of separation for burn injury hospitalisation, and with inclusion of multiple malignancies, 759 cancer notifications were included in the SIR analyses as independent observations.

In Scotland from 1983 to 2008, there were a total of 37 890 persons hospitalised for an index burn-related injury. After exclusion of those with a history of cancer prior to separation date or death during hospital admission for burn, a total of 37 506 patients were included in the analysis. There were 2005 patients with a first cancer notification after the date of separation for burn injury hospitalisation, and with inclusion of multiple malignancies, 2260 cancer notifications were included in the SIR analyses as independent observations. The characteristics of the Western Australia and Scotland cohorts are presented in [table 1](#).

The Western Australia cohort (1983–2008; combined gender) was followed for a total of 283 306 person-years, with a mean follow-up time of 12.3 years (range >0–25.9 years). The mean follow-up time for those with a cancer notification was 9.4 years (range >0–25.4 years) and for those with no cancer notification was 12.4 years (range >0–25.9 years). The Scotland cohort (1983–2008; combined gender) was followed for a total of 474 489 person-years, with a mean follow-up time of 12.6 years (range 0–26.0 years). The mean follow-up time for those with a cancer notification was 9.4 years (range >0–25.8 years) and for those with no cancer notification was 12.7 years (range >0–26 years).

For the Scottish cohort of burn-injury patients (combined gender), there was a marginal but significant difference (SIR, 95% CI 1.09, 1.05 to 1.10) in the overall risk of cancer for persons with a burn injury hospitalisation for the period 1983–2008, compared with the general population of Scotland. While a significant

**Table 1** Characteristics of burn injury patients included in analyses with no record of cancer prior to separation date of index burn admission, 1983–2008, by country

Characteristics	Western Australia N (%)	Scotland N (%)
Total number burn admissions*	22 705	3 537 506
Gender: male	15 481 (68.2)	23 896 (63.7)
Age at index admission (years)		
<15	8135 (35.8)	14 579 (38.9)
15–24	4364 (19.2)	4495 (12.0)
25–49	7147 (31.5)	9554 (25.5)
50–64	1736 (7.7)	4080 (10.9)
65+	1323 (5.8)	4798 (12.8)
Site of burn†		
Head and neck	6784 (15.4)	7592 (16.1)
Trunk	7553 (17.2)	8815 (21.0)
Hand, wrist, upper limb	15 801 (35.9)	6984 (14.8)
Hip, lower limb	11 798 (26.8)	9531 (3.4)
Eye	379 (0.9)	1087 (2.3)
Respiratory tract	212 (0.5)	163 (0.3)
Other internal organs	124 (0.3)	165 (0.3)
Multiple regions	656 (1.5)	3677 (7.8)
Unspecified region	694 (1.6)	858 (1.8)
Burn site depth†		
Erythema	8929 (20.9)	4815 (11.5)
Partial thickness	18 449 (41.9)	6302 (15.0)
Full thickness	7095 (16.1)	4924 (11.7)
Unspecified	9528 (21.7)	25 869 (61.7)
Calendar period of admission		
1983–1988	5431 (23.9)	11 507 (30.7)
1989–1993	4200 (18.5)	7876 (21.0)
1994–1998	4755 (20.9)	7130 (19.0)
1999–2003	4265 (18.9)	5980 (15.9)
2004–2008	4054 (17.9)	5013 (13.4)
Any comorbidity at index burn		
Yes	2798 (12.3)	7679 (20.5)

\*No previous record of cancer.

†Patients may have multiple burn sites per anatomical region per depth.

increase of 30% in cancer risk was estimated for females, there was no difference in cancer risk for males, when compared with the general population of Scotland (refer to [table 2](#)). For the subcohort of burn injury patients hospitalised during 1983–1988, the total observed number of cases of cancer (n=838) was statistically significantly lower than expected (n=953.4) with SIR (95% CI) of 0.88 (0.82 to 0.94), with males having a statistically significantly lower number of cases observed than expected. Refer to [table 2](#) for gender-specific SIRs for total (all sites) cancer for Scotland and Western Australian burn patients, hospitalised during 1983–1988 and 1983–2008.<sup>11</sup>

Female genital cancers were grouped due to the small number of observed cancers in subgroups in the Western Australian data and unstable SIR results. Statistically

**Table 2** SIRs and 95% CIs and observed and expected number for total (all sites) cancer in persons hospitalised for burn injury in Western Australia and Scotland, during the periods 1983–2008 and 1983–1988

	Western Australia*			Scotland		
	Combined SIR 95% CI‡ O:E	Male† SIR 95% CI O:E	Female† SIR 95% CI O:E	Combined SIR 95% CI O:E	Male† SIR 95% CI O:E	Female† SIR 95% CI O:E
Total cohort	0.97 (0.9 to 1.0)	0.9 (0.8 to 1.0)	1.1 (1.0 to 1.3)	1.09 (1.05 to 1.10)	0.96 (0.90 to 1.0)	1.3 (1.2 to 1.4)
1983–2008	759: 785.5	515: 569.5	244: 216.0	2260: 2075.9	1249: 1303.2	1011: 772.6
Subcohort	1.0 (0.9 to 1.1)	0.9 (0.8 to 1.0)	1.4 (1.1 to 1.7)	0.9 (0.8 to 0.9)	0.8 (0.7 to 0.9)	1.0 (0.9 to 1.2)
1983–1988	294: 294.9	190: 220.3	104: 74.6	838: 953.4	491: 614.3	347: 339.1

\*Western Australian comparison data Duke *et al.*<sup>11</sup>

†SIR (95% CI) adjusted for age.

‡SIR (95% CI) adjusted for age and gender.

O:E, observed:expected; SIR (95% CI), standardised incidence ratio (95% CI).

significant increases in observed genital (combined) cancers for female burn patients in Western Australia and Scotland were found. The increased breast cancer incidence was statistically significant among female burn survivors in Scotland. Statistically significant increases in cancer incidence for combined gender for Western Australia and Scottish data were observed for cancers of the buccal cavity, liver and respiratory tract. Refer to [table 3](#) for gender-specific and site-specific cancer SIRs. For each of these cancers, female burn survivors in Western Australia and Scotland had a higher incidence than males when compared with respective general population data. For the majority of site-specific cancers selected, female burn survivors in Western Australia and Scotland had a higher number of observed cancers than expected, with SIRs of similar magnitude. However, SIR results for Scottish data reached statistical significance, reflecting the larger population base and respective higher number of cancer notifications.

[Table 4](#) presents an SIR analyses of total (all sites) cancer risk repeated on age-restricted subcohorts, classified to reflect the reproductive age (<15; 15–49 and ≥50 years) at admission for burn injury. For males in both WA and Scotland, no statistically significant differences were found across the three age groups. For female burn survivors in Scotland, the observed number of total cancer (all sites) exceeded that expected for each of the three age groups, with statistically significant results observed for the age groups 15–49 and for those aged 50 years and older. In the Western Australia data, excess cancers were observed for those younger than 15 years and for those 50 years and older, with statistical significance reached for the older age group; for females aged 15–49 years at burn injury, no difference in observed and expected total (all sites) cancer was found.

## DISCUSSION

### Methodological issues

When population-level administrative data are used, data linkage minimises issues of selection and reporting bias, as well as loss to follow-up. Data quality of the Western

Australia and Scottish Cancer Registers<sup>17 19 20</sup> and hospital morbidity datasets<sup>24 25</sup> are assessed continually for accuracy and quality. Data of all index burn hospitalisations in Western Australia and Scotland from 1983 to 2008 were analysed with a follow-up time from discharge date, allowing for exclusion of prevalent cancers to support temporality of burn exposure and incident cancer. Cancer diagnoses from cancer registries in Western Australia and Scotland were independent of the record of burn injury in the respective hospital morbidity datasets. Minor burns treated in emergency departments were not included in the study. The burn patient cohorts under study are part of the respective reference populations, and as such, this may have a small diluting effect in the SIRs. Using parallel datasets from Western Australia and Scotland, with 2.2 million and 5.5 million populations, respectively, allowed examination of the consistency of results and trends across the populations.

The Western Australia hospital morbidity data records the principal diagnosis and up to 20 additional diagnosis fields, whereas the Scottish morbidity data include the principal diagnosis and five additional diagnosis fields. Consequent to the reduced number of additional diagnosis fields in the Scottish data, there was an absence of recorded supplementary total body surface area burned (TBSA%) data (ICD9 946; ICD10 T31) and a greater use of ICD codes for burns to multiple regions of the body (ICD9 946; ICD10 T29) rather than to individual anatomic burn sites, as reflected in [table 1](#). This limited SIR analyses restricted to more severe burns of TBSA 20% or greater and incident rate ratio analysis to examine the effects of severity of burn injury (burn depth and TBSA %). Previous SIR analyses of total (all sites) cancer risk in Western Australia showed similar trends in results for all burn patients (severe and non-severe).<sup>11</sup>

Although this study had a follow-up period of up to 26 years from the date of separation for admission for burn injury, the follow-up period for many patients may not have provided sufficient observation time to enable identification of all potential malignancies, given the long latency period for many cancers. Further burn injury research is planned with comparison cohorts (non-burn trauma, no injury), using incidence rate ratio



**Table 3** SIRs and 95% CIs and observed and expected numbers for selected types of cancer in persons hospitalised for burns in Western Australia and Scotland, 1983–2008

Cancer Site ICD-10*	Western Australia			Scotland		
	Combined SIR 95% CI† O:E	Male SIR 95% CI‡ O:E	Female SIR 95% CI‡ O:E	Combined SIR 95% CI‡ O:E	Male SIR 95% CI‡ O:E	Female SIR 95% CI‡ O:E
Buccal cavity	1.4 (1.03 to 1.9)	1.4 (1.0 to 1.9)	1.5 (0.7 to 3.2)	2.6 (2.2 to 3.1)	2.4 (1.9 to 2.9)	3.4 (2.5 to 4.8)
C00 to C14	45: 32.6	38: 28.1	7: 4.6	117: 45.0	83: 35.1	34: 9.9
Oesophagus	1.4 (0.9 to 2.4)	1.5 (0.9 to 2.6)	(0.3 to 4.5)	1.6 (1.3 to 2.0)	1.5 (1.1 to 1.9)	1.9 (1.3 to 2.7)
C15	15: 10.50	13: 8.7	2: 1.8	82: 51.4	53: 36.1	29: 15.3
Stomach	0.6 (0.3 to 1.1)	0.5 (0.2 to 1.1)	0.8 (0.3 to 2.6)	1.2 (0.9 to 1.5)	(0.8 to 1.5)	(0.9 to 1.9)
C16	10: 17.0	7: 13.4	3: 3.6	73: 63.2	5: 2.8	25: 19.5
Colorectal	0.7 (0.6 to 0.9)	0.7 (0.5 to 0.9)	0.9 (0.6 to 1.3)	1.2 (1.1 to 1.4)	(0.9 to 1.2)	(1.3 to 1.8)
C18 to C20	69: 96.3	45: 69.1	24: 27.2	268: 221.8	142: 140.5	125: 81.3
Liver	2.6 (1.6 to 4.0)	2.2 (1.3 to 3.7)	4.7 (2.0 to 11.4)	1.7 (1.2 to 2.5)	(1.1 to 2.5)	1.9 (1.0 to 3.7)
C22	19: 7.4	14: 6.3	5: 1.1	31: 18.0	22: 13.3	9: 4.7
Pancreas	0.7 (0.4 to 1.3)	0.9 (0.5 to 1.7)	0.4 (0.1 to 1.6)	1.1 (0.8 to 1.5)	1.5 (1.03 to 2.0)	0.6 (0.3 to 1.2)''
C25	11: 15.3	9: 10.4	2: 5.0	44: 39.6	34: 23.4	10: 16.2
Larynx	5.7 (0.9 to 3.3)	1.5 (0.7 to 3.0)	6.0 (1.5 to 24.1)	1.9 (1.4 to 2.5)	1.5 (1.1 to 2.2)''	4.2(2.3 to 7.7)
C32	10: 5.7	8: 5.4	2: 0.3	39: 21.1	28: 18.5	11: 2.6
Respiratory tract	1.4 (1.1 to 1.6)	1.3 (1.1 to 1.7)	1.4 (0.9 to 2.2)	1.5 (1.4 to 1.7)	1.3 (1.2 to 1.5)	1.9 (1.7 to 2.2)
C33 to C34	101: 74.8	79: 59.3	22: 15.4	448: 298.1	279: 210.4	169: 87.7
Skin—malignant melanoma	0.7 (0.6 to 0.9)	0.7 (0.6 to 1.0)	0.6 (0.4 to 1.0)	0.8 (0.6 to 1.1)''	0.7 (0.4 to 1.1)	(0.4 to 1.1)
C43	72: 102.0	57: 77.9	15: 24.1	38: 48.5	19: 28.4	19: 20.0
Breast''	(0.8 to 1.3)	1.3 (0.2 to 9.2)	1.0 (0.8 to 1.3)	1.7 (1.5 to 1.9)	0.7 (0.1 to 4.8)	(1.5 to 1.9)
C50	65: 62.4	1: 0.8	64: 61.7	271: 161.4	1: 1.5	270: 160.0
Female genital tract (combined)			1.4 (1.0 to 2.0)			1.7 (1.4 to 2.0)
C51 to C57			31: 26.7			114: 67.2
Male genital tract (combined)		0.9 (0.8 to 1.1)			1.1 (1.0 to 1.3)	
C60 to C63		141: 150.7			210: 192.6	
Prostate		0.8 (0.6 to 0.9)			1.1 (0.9 to 1.2)	
C61		102: 135.9			177: 165.5	
Kidney, Bladder, UT C64 to C68	0.5 (0.3 to 0.7)''	0.4 (0.2 to 0.7)	0.7 (0.3 to 1.7)	1.2 (1.0 to 1.4)	1.2 (1.0 to 1.4)	1.4 (1.0 to 1.9)
	17: 37.9	12: 30.9	5: 7.0	135: 110.9	96: 82.8	39: 28.0
Brain	1.2 (0.7 to 1.9)	1.0 (0.5 to 1.8)	1.7 (0.8 to 3.9)	1.5 (1.1 to 2.0)	1.4 (0.9 to 2.0)	(1.0 to 2.9)
C71	16: 13.9	10: 10.5	6: 3.5	39: 27.0	26: 19.2	13: 7.8
Lymphomas to all	(0.7 to 1.4)	0.8 (0.5 to 1.2)	1.7 (1.03 to 2.7)	1.1 (0.9 to 1.4)	1.1 (0.8 to 1.4)	1.2 (0.8 to 1.7)
	36: 35.5	20: 26.0	16: 9.6	75: 68.0	48: 45.0	27: 23.0
Myeloma/plasma	1.3 (0.7 to 2.3)	1.3 (0.7 to 2.6)	1.2 (0.4 to 3.7)	1.1 (0.7 to 1.6)	(0.6 to 1.7)	1.2 (0.6 to 2.2)
	11: 8.6	8: 6.1	3: 2.49	22: 21.0	13: 13.2	9: 7.8
Leukaemia's to all	1.1 (0.8 to 1.7)	1.1 (0.7 to 1.8)	1.2 (0.6 to 2.5)	1.3 (1.01 to 1.7)	(0.73 to 1.4)	(1.3 to 2.7)
	26: 22.9	19: 17.0	7: 6.0	63: 48.6	34: 33.1	29: 15.5

\*ICD-10: International Classification of Diseases V.10.

†SIR (95% CI) adjusted for age and gender.

‡SIR (95% CI) adjusted for age (95% CI).

O:E, Observed:expected; SIR (95% CI), standardised incidence ratio (95% CI).

**Table 4** SIRs and 95% CIs and observed and expected number for total (all sites) cancer incidence, for persons hospitalised for burns in Western Australia and Scotland, by age group, 1983–1988

Age at first burn years	SIR (95% CI) (observed: expected)		
	Combined gender*	Male†	Female‡
<15			
WA	1.17 (0.82 to 1.68) (30 : 25)	1.19 (0.77 to 1.84) (20 : 16)	1.15 (0.62 to 2.14) (10 : 8.6)
Scotland	0.94 (0.69 to 1.28) (41 : 43.69)	0.72 (0.47 to 1.12) (20 : 27.77)	1.32 (0.86 to 2.02) (21 : 15.92)
15–49			
WA	0.87 (0.77 to 0.99) (273 : 313)	0.87 (0.75 to 1.00) (197 : 226)	0.86 (0.69 to 1.1) (76 : 87)
Scotland	1.21 (1.12 to 1.31) (617 : 509.16)	1.04 (0.94 to 1.16) (345 : 331.68)	1.53 (1.36 to 1.73) (272 : 177.48)
≥50			
WA	1.02 (0.93 to 1.12) (456 : 446)	0.91 (0.82 to 1.02) (298 : 326)	1.32 (1.13 to 1.54) (158 : 120)
Scotland	1.05 (1.00 to 1.11) (1602 : 1523)	0.94 (0.88 to 1.00) (884 : 943.75)	1.23 (1.15 to 1.33) (718 : 579.25)

\*SIR (95% CI) adjusted for age and gender.

†SIR (95% CI) adjusted for age.

SIR (95% CI), standardised incidence ratio (95% CI); WA, Western Australia.

analyses to explore the patient (including lifestyle factors such as smoking and alcohol) and injury factors associated with the observed cancer risk.

### Findings

Analysis of the ISD Scotland data confirmed the results of our previous study: a statistically significant increase in total (all sites) cancer risk for female burn survivors with males experiencing no difference. The site-specific analyses clearly showed statistically significant increases in the number of observed cancers for combined gender in the Western Australia and Scottish data for the buccal cavity, larynx, liver, respiratory tract and oesophagus. There was also a general trend for increased cancer risk for a number of selected types of cancers for females and statistically significant increases in female genital cancers. Sub group analyses, defined crudely by reproductive age, did not elucidate any clear patterns of influence of oestrogen on cancer incidence, with female burn survivors in Scotland showing an increased risk across all age groups. For female burn survivors in WA, an increased risk for total (all sites) cancer was found for those younger than 15 (prepubescent) and 50 years and older (postmenopausal). The lack of gender difference for the subcohort of burn patients in Scotland hospitalised during 1983–1988 for total (all sites) cancer risk is difficult to explain. Possible reasons may include that females sustained less severe (<20% TBSA) burns during this period; had less comorbidities and/or had better lifestyle factors than females hospitalised for burns during the remainder of the study period.

The site-specific analyses showed that while statistically significant increases in female genital cancers were

found, there was also a general trend among female burn patients for excesses across a number of site-specific cancers examined, although these excesses did not always reach statistical significance, possibly due to small numbers. Statistically significant increases in the number of observed cancers for combined gender were found in the Western Australia and Scottish data for the buccal cavity, larynx, liver, respiratory tract and oesophagus. These results are similar to those found in a Danish study<sup>13</sup> and may be related to tobacco or alcohol use among this patient population. However, it would be expected that inhalation injury may also increase the cancer risk of the upper and lower respiratory tract, and in the case of the diagnosis of hepatocellular cancer in a young male (12 years of age) burn patient in Western Australia,<sup>10</sup> tobacco or alcohol use would be most unlikely attributable agents. Interestingly, the results of no increase in skin melanoma risk after burn injury in this study support findings of other population-based studies.<sup>12 13</sup>

An alternative explanation for this increased incidence in cancer postburn may lie in the significant impact a burn injury has on the immune system, or the sustained oxidative and metabolic stress that are integral to the injury response. While burn injury predominantly affects the skin, it has been shown to cause a significant depression of humoral and cell-mediated immunity,<sup>7 26 27</sup> sustained elevated levels of oxidative stress<sup>28 29</sup> and prolonged elevation of hypermetabolic and stress hormone levels.<sup>30 31</sup> These effects have been demonstrated to persist for up to 3 years postinjury and can lead to long-term systemic impacts on other organs of the body.<sup>1 2 32–36</sup> Severe burn injury has been demonstrated to induce endoplasmic reticulum (ER) stress, in

particular, in the liver.<sup>37</sup> ER stress is a stress-response that initially facilitates cell survival but can switch to a proapoptotic signal with prolonged stress.<sup>37–38</sup> However, it has also been shown that the ER stress response can become maladaptive, facilitating adaptation to hypoxic environments and promoting tumour growth.<sup>38–39</sup> It is plausible that the array of host responses combined with the impact of the injury, therefore, creates an environment of increased susceptibility to cancer.

In addition to the observed increase in some of the selected site-specific cancers, the data support evidence for a gender dimorphism (a systematic difference between individuals of different sex in the same species) in response to burn injury. After burn injury, gender has been shown to be an important factor with respect to poorer outcomes for mortality<sup>40–43</sup> and improved prognosis for multiple organ dysfunction syndrome,<sup>44</sup> and sepsis,<sup>45</sup> for females compared with males. Similar gender-based differences have also been reported in animal studies of burn injury.<sup>46–50</sup>

The impact of gender with respect to outcomes after burn injury is largely thought to stem from well-established differences in immune biology. There is a substantial volume of published literature to support a gender dimorphism in the immune response<sup>51–54</sup> and sepsis<sup>45</sup> following injury that have impacts on health and mortality.<sup>41–42</sup> The majority of these studies support a more efficient and effective innate and adaptive immune responses in females, leading to a rapid clearance of infectious organisms driven by tissue resident cell populations.<sup>55</sup> This 'advantageous' response reduces the risk of infection in females compared with males<sup>55–56</sup> but leads to elevated risk of autoimmune disease.<sup>57</sup> This dimorphism was thought to arise largely due to the impact of oestrogen on immune function.<sup>58–59</sup> However, recent papers have demonstrated that these differences are not completely ablated by ovariectomy (in animal models)<sup>55</sup> and others have shown that oestrogen can be deleterious to the immune response.<sup>60</sup> This suggests a role of other mediators, most likely expressed on the X chromosome, in the maintenance of the differential immune response.<sup>61–62</sup> The evidence for gender differences in the immune response, to thermal and other trauma, and its impact on outcomes is substantial. Here, the evidence of an increased cancer incidence in selected types of cancer after burn injury, with a greater effect in females, suggests the systemic immune response to burn injury may be a mediator of cancer susceptibility.

## CONCLUSION

Using population-based linked data of all burn patients in Western Australia and Scotland, consistent trends were found in excesses in cancer notifications for a range of selected site-specific cancers with an elevated and more widespread increase in female burn patients. Overall, however, the increased cancer risk affected a small proportion of the respective burn patient cohorts. More research is

required to understand the underlying mechanism(s) that may link burn injury to an increased risk of some cancers and why this is elevated in females, which may in turn enable identification of possible sites for intervention.

## Author affiliations

<sup>1</sup>Burn Injury Research Unit, School of Surgery, University of Western Australia, Crawley, Western Australia, Australia

<sup>2</sup>Population Health Research Network, Centre for Data Linkage, Curtin University, Perth, Western Australia, Australia

<sup>3</sup>Burns Service of Western Australia, Royal Perth Hospital and Princess Margaret Hospital, Perth, Western Australia, Australia

<sup>4</sup>Fiona Wood Foundation, Crawley, Western Australia, Australia

**Acknowledgements** The authors would like to thank the staff of the Health Information Linkage Branch for access to the Western Australian Data Linkage System and Scottish Record Linkage team for their assistance in obtaining the data and providing advice on aspects of coding. Furthermore, the authors would like to thank the WA Health Data Custodians for access to the core health datasets and the Western Australian Department of Health and Information Service Division (ISD) Scotland for their assistance and advice.

**Contributors** JMD and JB participated in planning, conduct and reporting. JB, MWF, SR and FMW participated in reporting. JMD and JB are the guarantors.

**Funding** Project data costs were supported by an Australian National Health and Medical Research Council (533502) research grant and a Raine Medical Research Foundation Priming grant (12/2013). JMD, Senior Research Fellowship, is supported by Woodside corporate sponsorship via the Fiona Wood Foundation.

**Competing interests** None.

**Ethics approval** Department of Health Western Australia Human Research Ethics Committee; National Health Service (NHS) National Services Scotland Privacy Advisory Committee; University of Western Australia Human Research Ethics Committee.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>

## REFERENCES

1. Anderson JR, Zorbas JS, Phillips JK, *et al.* Systemic decreases in cutaneous innervation after burn injury. *J Invest Dermatol* 2010;130:1948–51.
2. Jeschke MG, Gauglitz GG, Kulp GA, *et al.* Long-term persistence of the pathophysiologic response to severe burn injury. *PLoS ONE* 2011;6:e21245.
3. 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.
4. Deveci M, Sengezer M, Bozkurt M, *et al.* Comparison of lymphocyte populations in cutaneous and electrical burn patients: a clinical study. *Burns* 2000;26:229–32.
5. Heideman M, Bengtsson A. The immunologic response to thermal injury. *World J Surg* 1992;16:53–6.
6. McGill SN, Cartotto RC. Herpes simplex virus infection in a paediatric burn patient: case report and review. *Burns* 2000;26:194–9.
7. O'Sullivan ST, O'Connor TP. Immunosuppression following thermal injury: the pathogenesis of immunodysfunction. *Br J Plastic Surg* 1997;50:615–23.
8. Sjoberg T, Mzezewa S, Jonsson K, *et al.* Immune response in burn patients in relation to HIV infection and sepsis. *Burns* 2004;30:670–4.

9. Jeschke MG, Barrow RE, Herndon DN. Extended hypermetabolic response of the liver in severely burned pediatric patients. *Arch Surg* 2004;139:641–7.
10. Harper A, Rea S, Wood F. Hepatocellular carcinoma in a young survivor of major burns. *Burns* 2008;34:572–4.
11. Duke J, Rea S, Semmens J, et al. Burn injury and cancer risk: a state-wide longitudinal study. *Burns* 2011;38:340–7.
12. Lindelof B, Krynitiz B, Granath F, et al. Burn injuries and skin cancer: a population-based cohort study. *Acta Dermato-Venereol* 2008;88:20–2.
13. Mellekjaer L, Holmich LR, Gridley G, et al. Risks for skin and other cancers up to 25 years after burn injuries. *Epidemiology* 2006;17:668–73.
14. Holman CDJ, Bass AJ, Rouse IL, et al. Population-based linkage of health records in Western Australia: development of a health service research linked database. *Aust N Z J Public Health* 1999; 23:453–9.
15. Kendrick S, Clarke J. The Scottish Record Linkage System. *Health Bull* 1993;51:72–9.
16. 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.
17. Threlfall T, Thompson JR. *Cancer incidence and mortality in Western Australia, 2008. Statistical Series Number 87*. Perth: Department of Health, Western Australia, 2010.
18. ISD Scotland. ISD Scotland, better information, better decisions, better health, <http://www.isdscotland.org> (accessed 12 Aug 2013).
19. Brewster DH, Stockton D, Harvey J, et al. Reliability of cancer registration data in Scotland, 1997. *Eur J Cancer* 2002;38:414–17.
20. Brewster DH, Stockton DL. Ascertainment of breast cancer by the Scottish Cancer Registry: an assessment based on comparison with five independent breast cancer trials databases. *Breast* 2008;17:104–6.
21. Gordis L. *Epidemiology*, 2nd edn. Philadelphia: W.B. Saunders Company, 2000.
22. Verkasalo PK, Pukkala E, Kaprio J, et al. Magnetic fields of high voltage power lines and risk of cancer in Finnish adults: nationwide cohort study. *BMJ* 1996;313:1047–51.
23. Bailar J III, Ederer F. Significance factors for the ratio of a poisson variable to its expectation. *Biometrics* 1964;20:639–43.
24. ISD Scotland. *Assessment of SMR01 Data 2010–2011 Scotland Report May 2012*. Edinburgh: NHS National Services Scotland, 2012.
25. Department of Health Western Australia. Clinical information audit program—hospital activity report. *Operational Directive OD 0201/09*. Perth Department of Health WA, 2009.
26. Horgan PG, Mannick JA, Dubravec DB, et al. Effect of low dose recombinant interleukin 2 plus indomethacin on mortality after sepsis in a murine burn model. *Br J Surg* 1990;77:401–4.
27. Schmand JF, Ayala A, Chaudry IH. Effects of trauma, duration of hypotension, and resuscitation regimen on cellular immunity after hemorrhagic shock. *Crit Care Med* 1994;22:1076–83.
28. Liu DM, Sun BW, Sun ZW, et al. Suppression of inflammatory cytokine production and oxidative stress by CO-releasing molecules-liberated CO in the small intestine of thermally-injured mice. *Acta Pharmacol Sin* 2008;29:838–46.
29. Shupp JW, Nasabzadeh TJ, Rosenthal DS, et al. A review of the local pathophysiologic bases of burn wound progression. *J Burn Care Res* 2010;31:849–73.
30. Atiyeh BS, Gunn SWA, Dibo SA. Metabolic implications of severe burn injuries and their management: a systematic review of the literature. *World J Surg* 2008;32:1857–69.
31. Williams FN, Herndon DN, Jeschke MG. The hypermetabolic response to burn injury and interventions to modify this response. *Clin Plastic Surg* 2009;36:583–96.
32. Ananthkrishnan P, Cohen DB, Xu DZ, et al. Sex hormones modulate distant organ injury in both a trauma/hemorrhagic shock model and a burn model. *Surgery* 2005;137:56–65.
33. Dreschler S, Weixelbaumer K, Raeven P, et al. Relationship between age/gender-induced survival changes and the magnitude of inflammatory activation and organ dysfunction in post-traumatic sepsis. *PLoS ONE* 2012;7:e5147.
34. Borue X, Lee S, Grove J, et al. Bone marrow-derived cells contribute to epithelial engraftment during wound healing. *Am J Pathol* 2004;165:1767–72.
35. Fan Q, Yee CL, Ohyama M, et al. Bone marrow-derived keratinocytes are not detected in normal skin and only rarely detected in wounded skin in two different murine models. *Exp Hematol* 2006;34:672–9.
36. Harris RG, Herzog EL, Bruscia EM, et al. Lack of a fusion requirement for development of bone marrow-derived epithelia. *Science* 2004;305:90–3.
37. Jeschke MG, Finnerty CC, Herndon DN, et al. Severe injury is associated with insulin resistance, endoplasmic reticulum stress response, and unfolded protein response. *Ann Surg* 2012;255:370–8.
38. Verfaillie T, Garg AD, Agostinis P. Targeting ER stress induced apoptosis and inflammation in cancer. *Cancer Lett* 2013;332:249–64.
39. Wang S, Kaufman RJ. The impact of the unfolded protein response on human disease. *J Cell Biol* 2012;197:857–67.
40. George RL, McGwin G Jr, Schwacha MG, et al. The association between sex and mortality among burn patients as modified by age. *J Burn Care Rehabil* 2005;26:416–21.
41. Kerby JD, McGwin G Jr, George RL, et al. Sex differences in mortality after burn injury: results of analysis of the National Burn Repository of the American Burn Association. *J Burn Care Res* 2006;27:452–6.
42. McGwin G Jr, George RL, Cross JM, et al. Gender differences in mortality following burn injury. *Shock* 2002;18:311–15.
43. O'Keefe GE, Hunt JL, Purdue GF. An evaluation of risk factors for mortality after burn trauma and the identification of gender-dependent differences in outcomes. *J Am Coll Surg* 2001;192:153–60.
44. Frink M, Pape H-C, van Griensven M, et al. Influence of sex and age on mods and cytokines after multiple injuries. *Shock* 2007; 27:151–6.
45. Schroder J, Kahlke V, Book M, et al. Gender differences in sepsis: genetically determined? *Shock* 2000;14:307–10; discussion 10–3.
46. Gregory MS, Duffner LA, Faunce DE, et al. Estrogen mediates the sex difference in post-burn immunosuppression. *J Endocrinol* 2000;164:129–38.
47. Gregory MS, Faunce DE, Duffner LA, et al. Gender difference in cell-mediated immunity after thermal injury is mediated, in part, by elevated levels of interleukin-6. *J Leukoc Biol* 2000;67:319–26.
48. Kahlke V, Angele MK, Ayala A, et al. Immune dysfunction following trauma-haemorrhage: influence of gender and age. *Cytokine* 2000;12:69–77.
49. Kahlke V, Angele MK, Schwacha MG, et al. Reversal of sexual dimorphism in splenic T lymphocyte responses after trauma-hemorrhage with aging. *Am J Physiol Cell Physiol* 2000;278:C509–16.
50. Plackett TP, Gamelli RL, Kovacs EJ. Gender-based differences in cytokine production after burn injury: a role of interleukin-6. *J Am Coll Surg* 2010;210:73–8.
51. Croce MA, Fabian TC, Malhotra AK, et al. Does gender difference influence outcome? *J Trauma Injury Infect Crit Care* 2002;53:889–94.
52. Horton JW, White DJ, Maass DL. Gender-related differences in myocardial inflammatory and contractile responses to major burn trauma. *Am J Physiol Heart Circ Physiol* 2004;286:H202–13.
53. 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.
54. Verthelyi D. Sex hormones as immunomodulators in health and disease. *Int Immunopharmacol* 2001;1:983–93.
55. Scotland RS, Stables MJ, Madalli S, et al. Sex differences in resident immune cell phenotype underlie more efficient acute inflammatory responses in female mice. *Blood* 2011;118:5918–27.
56. Sperry JL, Nathens AB, Frankel HL, et al. Characterization of the gender dimorphism after injury and hemorrhagic shock: are hormonal differences responsible? *Crit Care Med* 2008;36:1838–45.
57. Ozcelik T. X chromosome inactivation and female predisposition to autoimmunity. *Clin Rev Allergy Immunol* 2008;34:348–51.
58. Nicol T, Vernon-Roberts B, Quantock DC. Effect of orchidectomy and ovariectomy on survival against lethal infections in mice. *Nature* 1966;211:1091–2.
59. Paavonen T. Hormonal regulation of immune responses. *Ann Med* 1994;26:255–8.
60. Rettew JA, Huet YM, Marriot I. Estrogens augment cell surface TLR4 expression on murine macrophages and regulate sepsis susceptibility in vivo. *Endocrinology* 2009;150:3877–84.
61. Libert C, Dejager L, Pinheiro I. The X chromosome in immune functions: when a chromosome makes the difference. *Nat Rev Immunol* 2010;10:594–604.
62. Pinheiro I, Dejager L, Libert C. X-chromosome-located microRNAs in immunity: might they explain male/female differences? The X chromosome-genomic context may affect X-located miRNAs and downstream signaling, thereby contributing to the enhanced immune response of females. *Bioessays* 2011;33:791–802.

## 6.7. Supporting Manuscript(s)

Capewell S, Kendrick S, **Boyd J**, Cohen G, Juszczak E, Clarke J. *Measuring outcomes: one month survival after acute myocardial infarction in Scotland*. Heart (1996)



MacIntyre K, Capewell S, Stewart S, Chalmers JWT, Boyd J, Finlayson A, Redpath A, Pell JP and McMurray JJV. ***Evidence of improving prognosis in heart failure: trends in case fatality in 66 547 patients hospitalized between 1986 and 1995.*** Circulation (2000)





MacIntyre K, Stewart S, Capewell S, Chalmers JWT, **Boyd J**, Finlayson A, Redpath A, Gilmour H. ***Gender and survival: a population-based study of 201,114 men and women following a first acute myocardial infarction.*** Journal of the American College of Cardiology (2001)



Capewell S, MacIntyre K, Stewart S, Chalmers JWT, **Boyd J**, Finlayson A, Redpath A, Pell J, McMurray JJV. ***Age, sex, and social trends in out-of-hospital cardiac deaths in Scotland 1986-95: a retrospective cohort study.*** The Lancet (2001)



## Chapter 7

---

### Conclusion

*“Finally, in conclusion, let me say just this.....”*

*Peter Sellers*

#### ***Professional Report(s):***

**Boyd JH**, Ferrante AM. Senate Select Committee on Health. *Improving access to and linkage between health dataset*. Public Hearing, December 2015

**Boyd JH**, Ferrante AM. Submission to the Productivity Commission: *Data Availability and Use*. July 2016



## 7.1. Conclusion

Government at all levels in Australia collect large amounts of administrative health and health related data about services for the community [163]. This data is gathered using agreed standards and definitions that provide a sound basis for quality and completeness of the various collections [150]. Good information collected consistently is essential for an efficient and effective healthcare system and for understanding health outcomes and service use [8-10].

Research using data linkage in Australia (and overseas) has already demonstrated its value in supporting policy decisions and improving public services [51, 164]. Data linkage provides information on whole populations and generates a more complete picture of the community than is possible using other research methods. It is also a very cost-effective research tool. Once the linkage infrastructure is in place, the cost of accessing linkable data becomes more affordable.

## 7.2. Solid infrastructure

The Population Health Research Network (PHRN), funded through the National Collaborative Research Infrastructure Strategy (NCRIS) [69], has been successful in developing linkage infrastructure that provides a platform to run large linkage projects (across state, national and cross-jurisdictional datasets) [146, 147, 165]. Australian data linkage infrastructure is recognised internationally for its high level of accuracy (linkage quality standards) and innovative technologies/methodologies. The PHRN infrastructure has enhanced Australia's ability to conduct high quality, internationally competitive research. The challenge is to continue to realise the potential of the infrastructure and develop a sustainable and effective model for the future.

This PhD has built upon collaborations with the PHRN to gain a deeper understanding of the current status and issues relating to record linkage and linkage quality, and to assist in identifying and addressing challenges and bottlenecks across the network. Using partnership and relationships with both university and government researchers (nationally and internationally) to improve the quality of linkage methods, systems and operations for the Australian research community. The PhD research programme has helped enhance and operationalise national record linkage protocols and provided opportunities to extend population research using linked data and to foster new collaborations.

The development of a robust linkage model with innovative technology has helped improve matching accuracy and quality of the links provided to researchers. Information on linkage quality allows researchers to assess/address any bias in the study design (e.g. if data is coming from different systems, are the data and linkage results consistent?) and allow adjustment to statistical confidence levels in the interpretation of results. Little work had been carried out on developing standard quality processes for assessing, improving and reporting on the quality of linkage outputs [119, 125]. This project has helped fill this information gap.

### **7.3. A career in record linkage**

I have been very fortunate to be at the 'cutting edge' of record linkage development throughout my professional career. Following graduation, my first job was in the newly formed record linkage unit within the information headquarters of the National Health Service (NHS) in Scotland. Throughout my various roles within the NHS, record linkage was always a primary focus.

The record linkage unit provided an opportunity to work with a variety of national data collections, disease registries and research datasets for national statistics, epidemiology and medical research. I also gained experience working with a wide range of stakeholders (including clinicians, health and education professionals, research bodies and University departments) to design, scope, analyse and interpret information for specific research projects and audits, negotiating with clients as required.

As part of the Year 2000 system redevelopment, my role focused on the development and delivery of a production linkage system to provide monthly updates to the Scottish linked morbidity and mortality file. The resulting record linkage service, added value to existing national datasets by routinely relating together health and healthcare activity for the same individual.

The Scottish linkage system is a unique resource, used to produce national epidemiological and management information to assist in monitoring and evaluating NHS resources and performance. The system provides accurate up-to-date linked information which is responsive to customer needs.

### **7.4. An unexpected journey**

Having led the record linkage work program in Scotland, I was approached by Curtin University in 2008 to establish national linkage infrastructure for Australia. Although there



were parallels between the requirements and challenges in Scotland and Australia, this has been a unique opportunity to research, develop and refine new record linkage methods, models, tools and algorithms.

This PhD programme has provided me with an opportunity to explore areas which have previously been barriers or limitations to the record linkage process. It has equipped me with skills and experience to further develop my interest in data linkage and the analysis of large complex datasets.

Research for the PhD addressed technical and methodological challenges associated with large-scale data linkage. This has led to the creation of enterprise level linkage infrastructure that provides a platform for undertaking large linkage projects while meeting the requirement to provide a secure and controlled environment for working with sensitive data. The infrastructure has been designed to scale as dataset size and demand for national linked data increases. This is critical as national data linkage in Australia moves towards the inclusion of new datasets.

With limited experience of working in an academic environment, the PhD programme (with mentoring from the Chair in Health Innovation, Population Health) has extended my expertise in developing grant applications with comprehensive research protocols for ongoing research.

During the PhD process, and the establishment of PHRN national services, I had the opportunity to spend time at the Australian Institute of Health and Welfare (AIHW) to lead the Institute's transition to become an accredited Commonwealth Integrating Authority. This provided an opportunity to work within the structures and processes of the Commonwealth Government. From developing innovative technology to providing advice, training and support; the PhD programme has facilitated the establishment of linkage services across Australia.

I believe that the research undertaken as part of the PhD has strengthened Australia's record linkage infrastructure and provided opportunities for research collaborations with other national and international data linkage centres. The research has also expanded and improved the scope and functionality of traditional linkage systems to incorporate important new data sources and develop new data linkages using cutting edge technologies and international intelligence in this field.

## 7.5. A vision for the future

Having been involved in record linkage for many years, I believe the main challenges to linkage in Australia (and around the world) are not related to the technical aspects of integrating data. They are often centred on more practical elements of data sharing, especially around the release of personal information across organisations. Obstructions to data sharing can be a result of legal or legislative barriers but are more often related to administrative 'red tape' or the disposition of data owners. These barriers can often be dismantled by mitigating risks associated with the data sharing process (through careful planning, secure protocols and legal agreements). However, government organisations often see research using administrative data as a threat in terms of their operation and expertise. Avoiding data sharing for data linkage removes any challenge to them and it is difficult for individuals outside government to understand or interface with their processes.

In recognition of these issues, the Senate Select Committee on Health held a public hearing into *Improving access to and linkage between health datasets* and the Productivity Commission (*Productivity Commission Act 1998*) has undertaken an inquiry into the *Benefits and costs of options for increasing availability of and improving the use of public and private sector data by individuals and organisations*. Both inquiries received a number of submissions (including responses prepared by the Centre for Data Linkage for Curtin University) outlining successes and the challenges with data sharing in an era where the capacity to collect, store and analyse data is expanding in an unprecedented way.

The resulting report prepared by the Productivity Commission (Data Availability and Use – currently draft) proposes a new legal and policy framework to allow public and private sector data to flow. The key elements of the framework focus on:

- Giving individuals more control over data held on them;
- Enabling broad access to datasets (public and private sector) that are of national interest;
- Increasing the usefulness of publicly funded identifiable data amongst trusted users; and
- Creating a culture in which non-personal and non-confidential data gets released by default for widespread use.

The recommendations in the Productivity Commission Report provide good foundations to build future data sharing models. However the current version still focuses on data flowing to

government agencies and it would be good to see some balance to the partnership between public and private sector organisations. However, I can see this stalemate around data sharing is being challenged through these open data policies and a need for government to work with private industry to maximise information available on community needs. These stakeholders need to make the most of data by integrating their data resources. It is also clear that those linkage systems that can use encryption and other advanced techniques to mask data prior to linkage will make the process of data sharing and linkage easier.

I believe that mainstream Privacy Preserving Record Linkage (PPRL) will be a major development internationally over the next five years. Aimed at reducing the need to use identifiers in their raw form for data linkage, PPRL systems will be adapted and scaled with sufficient quality to augment or potentially replace current approaches to data linkage. This will open new and innovative opportunities for the research community.

Curtin University is at the forefront of developing and appropriately adapting/scaling this technology for operational use in real-world settings [166]. As a major player in international research collaboration with University of Duisburg-Essen (Germany), Population Data BC (British Columbia), The Institute for Clinical Evaluative Sciences (ICES), PolicyWise (Edmonton), Swansea University (Wales) and Bristol University (England); we have been recognised as experts in this area. This international data linkage collaboration is well positioned to provide a platform to deliver world class data linkage and data analytics products and services.

The challenge will be to work with existing linkage experts to develop PPRL systems and models that compliment current linkage practice. It is clear that computational power and increasing sophistication in encryption techniques are removing the barriers to operationalising this technology. Given the sensitivity of the information involved, and the growing desire to link a broader number of datasets, any risk mitigation will be valued both by data owners and the public.

International data linkage collaboration will create opportunities for innovative knowledge exchange. This will extend the already successful collaboration demonstrated through the privacy preserving record linkage project and joint research on this linkage method which has been recognised internationally.

To complete the picture of advanced 'Big Data' analytics, the focus of infrastructure research is moving to harness data linkage and data analytic products and services [6]. Providing secure and efficient access to large integrated data within a powerful computing environment will unlock the potential within these data resources. In my view, the future of record linkage is connected to developments around analytical capabilities.

The thesis has shown the potential in record linkage technologies but more work is required. As this research develops, the resulting infrastructure and software will be both innovative and effective, reinforcing Australia's position as a world leader in the establishment and operation of record linkage for government and university research.

## 7.6. Professional Report(s)

**Boyd JH, Ferrante AM. *Senate Select Committee on Health. Improving access to and linkage between health datasets.* Senate Public Hearing. December 2015**



**Senate Select Committee on Health**  
*Improving access to and linkage between health datasets*

Submission from  
PHRN Centre for Data Linkage  
Curtin University

December 2015

This page has been left blank intentionally.



# Submission to Senate Select Committee on Health

---

## Current status of collection, linkage and access to health datasets

### Good data with great potential

The Australian Government, along with its state-based counterparts, collect large amounts of administrative health and health related data. The data is of good quality and provides a strong foundation for exploring and understanding health outcomes and service use. Good information is essential for an efficient and effective healthcare system and to ensure positive outcomes for individuals.

Research using data linkage in Australia and overseas has demonstrated its value in supporting policy and improving public services. “Data linkage” is a way of bringing data together and provides information on whole populations generating a more complete picture of the community than is possible using other research methods. It is also a very cost-effective research tool.

The benefits of data sharing have been shown to significantly improve research skills and analytical tools for complex “linked data”, enabling new research that enhances the delivery of public services. University based researchers in Australia are recognised as world leaders in the use of linked data for research. The demand for linked data from this sector is strong and growing.

Access to health datasets held at state/territory level has improved markedly since the establishment of the Population Health Research Network (PHRN). Mechanisms are in place in all Australian states and territories enabling access to data for research of national and international significance, and data “is flowing” in many state jurisdictions. *However, access to Commonwealth-held datasets remains an issue and the timeliness and burden of approval processes is also problematic. The Australian Government continues to have a poor track record in allowing researchers to access data or to link datasets.* It is important that this data is unlocked to allow research that can benefit the whole community.

### Solid infrastructure and information governance

Significant investment in data linkage infrastructure has occurred in Australia. Mature systems in WA and NSW have been supplemented by new infrastructure in other states and territories. The PHRN – an initiative funded through the National Collaborative Research Infrastructure Strategy (NCRIS) - has significantly expanded linkage infrastructure and provided national and cross-jurisdictional capabilities. PHRN “nodes” are characterised by high standards of information governance and data security to ensure privacy protection and confidentiality. Australian data linkage infrastructure is recognised internationally for its high level of accuracy (linkage quality standards) and innovative technologies/methodologies. The PHRN infrastructure has enhanced Australia’s ability to conduct high quality, internationally competitive research.

The Australian Government has also developed a data linkage/integration framework which incorporates data governance requirements and protection of patient privacy and confidentiality from existing government and PHRN processes and controls. Data linkage/integration involving

Commonwealth data can be carried out for statistical or research purposes by appointed Integrating Authorities. Integrating Authorities must ensure datasets are managed in a way that gives the community and businesses confidence that no individual or organisation is likely to be identified.

The development of the eHealth Record Systems (My Health Record) through the National E-Health Transition Authority (NEHTA) also provides opportunities for secondary use of health data for government and university research through data sharing and linkage with other information sources.

*In short, with innovative large scale infrastructure now in place for national and cross-jurisdictional linkage, the challenges are currently in negotiating information access and flows with state and Commonwealth data custodians.*

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 national or “Commonwealth” level), the impact of Commonwealth funding, serving planning and health outcomes can only be achieved through efficient cross-jurisdictional and national infrastructure. Experiences from other countries demonstrate the need to harness and harmonise the power and experience of linkage services and systems to improve the efficiency and quality within overall data linkage infrastructure.

A challenge in Australia is to realise the potential of the infrastructure currently available across government and university sectors through compatible, sustainable and effective models which can maximise the capacity across all these systems.

## **International developments**

Data linkage in the United Kingdom is undergoing a significant expansion of capabilities. Charities, Research Councils (the Medical Research Council and the Economic and Social Research Council), Government and other bodies have invested over £200million in the new Farr Institute - a collaborative ‘partnership model’ between government and the university sectors. The aim of the Farr Institute is to provide an integrated research platform for health and other Government sectors. Major centres are located in London, Dundee, Manchester and Swansea and link research in 19 universities across the UK and Northern Ireland.

The Farr Institute supports safe use of patient and research data for medical research across all diseases in the UK. Its research supports innovation in the public sector and industry leading to advances in preventative medicine, improvements in healthcare and better development of commercial drugs and diagnostics. The Farr Institute will also provide new insights into the understanding of causes of ill health which in turn will guide new biomedical research discovery. In preparation for these national developments, data linkage experts from Australia have provided advice and support to various Farr Institute nodes.

Legal, administrative and technical issues across the world have impacted on the ability to undertake linkage of particular datasets. New record linkage techniques, collectively referred to as privacy-preserving record linkage, significantly reduce privacy risks as they operate on de-identified information and do not require the release of personal identifiers. Researchers from Australia, Germany, Canada and the United Kingdom are developing software that implements Privacy

Preserving Record Linkage (PPRL) for use in operational record linkage settings. Adoption and application of these methods would increase capabilities and enable linked research opportunities as additional datasets are made available through a PPRL framework.

With significant international investment in data linkage and 'Big Data' science (supporting a push for open government) in the United Kingdom and Canada, long term funding of data linkage infrastructure in Australia is required to avoid losing the competitive advantage that Australia has gained in the international data linkage arena and in the fields of research that use linked data.

## **Challenges Faced**

Government and University departments in Australia understand that linked administrative data can provide an unparalleled resource for the monitoring and evaluation of services. However, for a number of reasons, these data have not been accessible to researchers. An additional barrier in Australia is that health data are collected by different levels of government – thus not all available through any one authority.

### **Barriers to data access**

The methods and techniques around data linkage in Australia are well established and the new developments (exploiting advances in technology) have the potential to improve timeliness and efficiency. The Commonwealth data linkage framework needs to build on existing national and international collaborations, infrastructure and skills. Leveraging these developments will fast track the research and policy development programme.

The main challenges or barriers to be resolved in realising the potential of the data linkage framework for Commonwealth health and health related data in Australia include:

- The distributed nature of health care responsibilities, coupled with a federated legal system, means that any long-term solution requires cooperation between state and Commonwealth stakeholders;
- Authorising environments - it is time-consuming to establish projects in terms of approvals and governance arrangements. Establishing transparent and consistent procedures that manage/streamline all the processes involved in executing a linkage project (end to end arrangements) would ensure effective and efficient data linkages. Transparent and consistent processes would reduce uncertainty around ethics, privacy and data custodian constraints;
- Legislation – many of the significant Commonwealth and state datasets are subject to specific legislation that define the conditions of data release and/or use. The extent of this type of legislation and its complexity creates difficulties of interpretation with regard to the release of data for linkage projects;
- Security, ethics and privacy – in addition to legal requirements, access to many Commonwealth and State health datasets are subject to privacy and ethical review. The processes necessary to address privacy and confidentiality concerns are not always transparent. It should be clear what governance process, protocols and standards are required to enable safe and secure access to linked data. In addition, the requirement for multiple ethics approvals (often in

different application forms) adds additional layers of bureaucracy within the project approval process;

- Operational efficiencies – The “quantity of data” emerging from electronic health collections also poses challenges (i.e. Big Data). Increasing demand on data linkage services also puts significant pressure on infrastructure to deliver in a timely fashion;
- Capacity – at present, the operations required to fulfil a data request can pose a substantial burden on organisational resources. Infrastructure enabling data linkage needs to be scalable, fast and efficient to ensure timely responses to important policy and research questions;
- Expertise – data linkage requires expertise in three broad areas – knowledge of the datasets available for linkage along with their characteristics and limitations, skills in linkage methods and skills in using/analysing linked information. By itself, a basic-level ability to use available linkage software is insufficient, because correct interpretation of linked datasets depends on an understanding of the structure and content of, and variation within the component datasets;
- The funding environment - The PHRN represents a major co-investment by the Commonwealth Government and PHRN partners in national data linkage infrastructure. However, the current funding model is inadequate (time-frames too short; uncertainties high) which makes operation, maintenance and support of the infrastructure difficult and innovation virtually impossible. Without long term funding the infrastructure will be unable to realise its full potential.

### **Proposals for Reform**

Under the data linkage/integration model adopted in Australia, Data Custodians play a pivotal role in granting approval and releasing data for linkage projects. These processes are numerous, lengthy and frequently subject to change. Data Custodians need to develop more effective models to support linkage-based research using health and health related data. Strategies and processes that reduce burden and improve turnaround times for linkage projects without impacting on costs are urgently required.

High level endorsement of frameworks (at ministerial level) and ‘whole of government’ approach would provide clarity within government and reduce the tendency for ‘defensive decision-making’ by Data Custodians.

Efficiencies could also be gained through:

- Cooperation: Development and endorsement of agreed principles/statements asserting value of data linkage for public benefit and supporting the release of data for data linkage;
- Streamlined access: Creating a streamlined and consistent application and approval processes for data linkage projects (especially for complex national/cross-jurisdictional projects using health and health related data). At present approvals processes are numerous and lengthy. Developing a national, co-ordinated approach to ethics applications and approvals is required to expedite access (simplify the process; reciprocal/mutual recognition). Moreover, national data linkage services could be provided through a “one-stop” shop;

- Enhance data flows: Exploring and implementing effective methods of enabling data flow (especially for complex, multi-dataset, multi-agency national or cross-jurisdictional projects). Including agreed data flows which provide comprehensive security and make collaborations between researchers easier and more efficient;
- Transparency and public accountability: Clear processes around assessments (for example, the balance of public good against the privacy imposition and risks to confidentiality) are essential as is public accountability. Accountability mechanisms could include the creation of an independent auditing or oversight body, community representation on steering committees or an advisory committee;
- Move from project to ongoing linkage: Linkage efficiency and quality could be improved by changing the current national data integration models to include routine/ongoing linkage of data and preservation of linkage maps;
- National data linkage services provided through a “one-stop” shop. Consider partnered or integrated service delivery to streamline and simplify current approaches to accessing linked data for national or cross-jurisdictional projects;
- Interoperability: Australia needs to ensure that data linkage infrastructure and technologies are interoperable and responsive to environmental changes around legislation, information technology, security and privacy. Common platforms allow the transfer of expertise, learning and skills between government and university teams. This would not only expand the capacity for national data linkage and reduce training costs associated with specialist linkage skills.

## Summary

Given the federated nature of health care service delivery in Australia, health research needs timely access to linked data from both State and Commonwealth Governments to answer questions regarding effectiveness and value for money in the health service. At present, the main barriers to linked health research and data sharing are around access and approvals. Defined processes and high level support can overcome these issues. Development of integrated project management and/or workflow systems or interfaces with existing systems such as the new National Ethics Application System will streamline some of these processes. Developing a transparent consistent approvals framework will help avoid duplication of effort and expedite research.

Increasing demand on data linkage services will put significant pressure on infrastructure to deliver in a timely fashion. The Commonwealth, in particular, needs to look at mechanisms which will ensure the timely delivery of data, particularly as the number, size and complexity of linkage research projects increase. Continuous improvement will also be essential. The infrastructure must continue to identify and implement new technologies to improve the efficiency of data linkage in Australia.

With linkage facilities in place across Australia (in States, Commonwealth departments and the PHRN), there are opportunities to leverage expertise and best practices. Development of interoperable system will increase capacity and speed of data linkage processes. Partnerships between national data linkage activities will drive innovation and efficiency within linkage systems.

Access to state and Commonwealth health data and a widening range of cross-sectoral data will increase the user-base of the infrastructure and enhance research outputs. Increased demand for national/cross jurisdictional linkage will provide economies of scale. The operational efficiency will increase with scale, leading to lower variable cost.

The Commonwealth Government should recognise that its own infrastructure is of international significance. This should be promoted, as well as the research that flows from having the infrastructure. The development of responsive, agile and efficient infrastructure which can assess and manage privacy risks will enable research that can improve health and enhance the delivery of healthcare and health related services.

**Boyd JH, Ferrante AM. *Submission to the Productivity Commission: Data Availability and Use.* July 2016**





# **Productivity Commission**

## ***Data Availability and Use***

Submission from  
Centre for Data Linkage  
Curtin University

July 2016

This page has been left blank intentionally.

# Submission to the Productivity Commission

---

## Data Availability and Use

Collection, storage and management of data have come a long way in the last two decades with significant developments in technology in terms of power and capacity. The volume of digital data is increasing exponentially, providing more available data for research. This increase in information also includes routinely collected administrative data which provides the building blocks for critical analysis used to shape policy, evaluate performance and improving public services.

Private and public organisations collect significant amounts of data on their business processes and services. Analysing these individual and aggregate records provide an opportunity to improve knowledge of the environment. Data manipulation and analytics can unlock the potential within the data and combining data from different sources can often provide a better understanding of the ‘big picture’.

“Data linkage” is a way of bringing data together to provide information on the whole population, generating a more complete picture of the community than is possible using other research methods. It is also a very cost-effective research tool.

The Centre for Data Linkage (CDL) is a national leader in the research and development of technology to safely and securely maximise the value of data available for research. The CDL was established in 2009 within Curtin under the National Collaborative Research Infrastructure Strategy (NCRIS) and is a member of the Population Health Research Network (PHRN). The focus was to develop and implement secure, state-of-the-art national infrastructure to enable cross-jurisdictional data linkage for research. As part of the project, the CDL has undertaken research into both technical and methodological aspects of data linkage to improve performance and capacity. These have been incorporated into the design and construction of efficient linkage infrastructure that maintains data integrity and security. The CDL has also provided technical advice and support to linkage units across Australia (PHRN). This support and guidance includes organising and running technical forums to discuss and share common challenges in designing, building and operating data linkage infrastructure. Since 2009, the CDL has:

- Designed, built and operated a secure environment to host the CDL data linkage infrastructure (including scalable Cross Jurisdictional linkage capabilities across Australia (1-4));
- Undertaken research into linkage systems, methods and models (including evaluation of linkage products and systems; development of Quality Assurance tools; a review of, and research into, data cleaning and standardisation; a review of privacy preserving data linkage techniques and a number of other on-going research projects). The quality of that research is evidenced through numerous publications in peer-reviewed journals (see Attachment A for list of publications);
- Adopted innovative approaches to data linkage functionality, performance, algorithms (including scalability in matching algorithms), parallel processing and database optimisation. The CDL has developed data linkage infrastructure that is reflective of

contemporary standards in IT and incorporates latest technologies in data linkage. These include, for example, a large, scalable linkage system with concurrent processing; multi-threading; differing grouping strategies; the ability to undertake project and enduring linkages; Privacy Preserving Record Linkage (PPRL) options; and

- Provided technical advice and assistance – includes direct assistance, customised training of technical staff, information sharing sessions with other PHRN members, publication and broader distribution of reports, promotion and hosting of PHRN Technical Forums. Technical leadership and innovation are evidenced through a variety of published articles on data linkage methods (5-10).

## **Questions on high value public sector data**

*What public sector datasets should be considered high-value data to the: business sector; research sector; academics; or the broader community?*

University based researchers in Australia are recognised as world leaders in the use of public sector data for research. Health, education and criminal justice datasets have been widely used to gain a better understanding of systems and services. Using a whole system approach recognises that many interacting factors can influence individual parts of the system and that solutions to problems have to be developed taking these factors and interactions into account. The demand for data from these sectors is strong and growing.

Research using routine administrative data in Australia and overseas has demonstrated its value, access to health, education and criminal justice datasets are crucial in supporting policy and improving public services.

*What characteristics define high-value datasets?*

Many of the administrative datasets are collected by government departments and other organisations to measure and monitor operations during the delivery of a service. The high-value characteristics of administrative data include coverage of the whole population, which allows analysis of small group and vulnerable, collection quality standards (with metadata) and long-term analysis to a level of detail not permitted by sample surveys. The use of administrative data provides a cost effective and efficient method for population based research and avoids imposing a further burden on respondents.

*What benefits would the community derive from increasing the availability and use of public sector data?*

The benefits of data sharing have been shown to improve research skills and analytical tools significantly for complex integrated data, enabling new research that enhances the delivery of public services. Access to high quality information is essential for efficient and effective government systems and services.

## **Questions on collection and release of public sector data**

*What are the main factors currently stopping government agencies from making their data available?*

Government and University departments in Australia understand that administrative data can provide an unparalleled resource for the monitoring and evaluation of services.

However, for a number of reasons, these data have not been accessible to researchers. An additional barrier in Australia is that health data are collected by different levels of government – thus not all available through any one authority.

## **Barriers to data access**

The main challenges or barriers to be resolved in realising the potential health and health related data in Australia include:

- The distributed nature of health care responsibilities, coupled with a federated legal system, means that any long-term solution requires cooperation between State and Commonwealth stakeholders;
- Authorising environments - it is time-consuming to establish projects in terms of approvals and governance arrangements. Establishing transparent and consistent procedures that manage/streamline all the processes involved in data access would ensure effective and efficient use of information. Transparent and consistent processes would reduce uncertainty around ethics, privacy and data custodian constraints;
- Legislation – many of the significant Commonwealth and State datasets are subject to specific legislation that defines the conditions of data release and/or use. The extent of this type of legislation and its complexity creates difficulties of interpretation with regard to the release of data for research projects;
- Operational efficiencies – The “quantity of data” emerging from electronic health collections also poses challenges (i.e. Big Data). Increasing demand on data linkage services also puts significant pressure on infrastructure to deliver in a timely fashion;
- Capacity – at present, the operations required to fulfil a data request can pose a substantial burden on organisational resources. Infrastructure enabling data linkage needs to be scalable, fast and efficient to ensure timely responses to important policy and research questions;
- Expertise – data linkage requires expertise in three broad areas: knowledge of the datasets available for linkage along with their characteristics and limitations, skills in linkage methods and skills in using/analysing linked information. By itself, a basic-level ability to use available linkage software is insufficient, because correct interpretation of linked datasets depends on an understanding of the structure and content of, and variation within the component datasets;
- The funding environment - The PHRN represents a major co-investment by the Commonwealth Government and PHRN partners in national data linkage infrastructure. However, the current funding model is inadequate (time-frames too short; uncertainties high) which makes operation, maintenance and support of the infrastructure difficult and innovation virtually impossible. Without long term funding the infrastructure will be unable to realise its full potential.

*How could governments use their own data collections more efficiently and effectively?*

Efficiencies could be gained through:

- Cooperation: Development and endorsement of agreed principles/statements asserting value of data for public benefit and supporting the release of data for research;

- Enhance data flows: Exploring and implementing effective methods of enabling data flow (especially for complex, multi-dataset, multi-agency national or cross-jurisdictional projects). Including agreed data flows which provide comprehensive security and make collaborations between researchers easier and more efficient;
- The development of the eHealth Record Systems (My Health Record) through the National E-Health Transition Authority (NEHTA) also provides opportunities for secondary use of health data for government and university research through data sharing and linkage with other information sources;
- Interoperability: Australia needs to ensure that infrastructure and technologies are interoperable and responsive to environmental changes around legislation, information technology, security and privacy. Common platforms allow the transfer of expertise, learning and skills between government and university teams.

*Should the collection, sharing and release of public sector data be standardised? What would be the benefits and costs of standardising? What would standards that are 'fit for purpose' look like?*

- Streamlined access: Creating a streamlined and consistent application and approval processes for research projects (especially for complex national/cross-jurisdictional projects using health and health related data). At present, approvals processes are numerous and lengthy. Developing a national, co-ordinated approach to ethics applications and approvals is required to expedite access (simplify the process; reciprocal/mutual recognition);
- Transparency and public accountability: Clear processes around assessments (for example, the balance of public good against the privacy imposition and risks to confidentiality) are essential as is public accountability. Accountability mechanisms could include the creation of an independent auditing or oversight body, community representation on steering committees or an advisory committee.

*What specific government initiatives (whether Australian Government, state, territory or local government, or overseas jurisdictions) have been particularly effective in improving data access and use?*

Unfortunately, it seems that all government agencies (national and international) seem unable to share unit data for research. Progressive policies, with suitable safeguards, around data sharing for research are required to maximise the value of collected data.

## **Questions on data linkage**

The methods and techniques around data linkage in Australia are well established, and the new developments (exploiting advances in technology) have the potential to improve timeliness and efficiency. Leveraging these developments will fast track the research and policy making programme.

*Which datasets, if linked or coordinated across public sector agencies, would be of high value to the community, and how would they be used?*

Health, education and criminal justice datasets provide a stable platform for both government and university research teams to gain a better understanding of population interactions with systems and services.

*Which rules, regulations or policies create unnecessary or excessive barriers to linking datasets?*

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 national or “Commonwealth” level), the impact of Commonwealth funding, serving planning and health outcomes can only be achieved through efficient cross-jurisdictional and national infrastructure. Experiences from other countries demonstrate the need to harness and harmonise the power and experience of linkage services and systems to improve the efficiency and quality within overall data linkage infrastructure.

- Authorising environments - it is time-consuming to establish projects in terms of approvals and governance arrangements. Establishing transparent and consistent procedures that manage/streamline all the processes involved in executing a linkage project (end to end arrangements) would ensure effective and efficient data linkages. Transparent and consistent processes would reduce uncertainty around ethics, privacy and data custodian constraints;
- Legislation – many of the significant state and commonwealth datasets are subject to specific legislation that defines the conditions of data release and/or use. The extent of this type of legislation and its complexity creates difficulties of interpretation with regard to the release of data for linkage projects;

*How can Australia’s government agencies improve their sharing and linking of public sector data? What lessons or examples from overseas should be considered?*

Data linkage in the United Kingdom is undergoing a significant expansion of capabilities. Charities, Research Councils (the Medical Research Council and the Economic and Social Research Council), Government and other bodies have invested over £200million in the new Farr Institute - a collaborative ‘partnership model’ between government and the university sectors. The aim of the Farr Institute is to provide an integrated research platform for health and other Government sectors. Major centres are located in London, Dundee, Manchester and Swansea and link research in 19 universities across the UK and Northern Ireland.

The Farr Institute supports safe use of patient and research data for medical research across all diseases in the UK. Its research supports innovation in the public sector and industry leading to advances in preventative medicine, improvements in healthcare and better development of commercial drugs and diagnostics. The Farr Institute will also provide new insights into the understanding of causes of ill health which in turn will guide new biomedical research discovery. In preparation for these national developments, data linkage experts from Australia have provided advice and support to various Farr Institute nodes.

Legal, administrative and technical issues across the world have impacted on the ability to undertake linkage of particular datasets. New record linkage techniques, collectively referred to as privacy-preserving record linkage, significantly reduce privacy risks as they operate on de-identified information and do not require the release of personal identifiers. Researchers from Australia, Germany, Canada and the United Kingdom are developing

software that implements Privacy Preserving Record Linkage (PPRL) for use in operational record linkage settings. Adoption and application of these methods would increase capabilities and enable linked research opportunities as additional datasets are made available through a PPRL framework.

With significant international investment in data linkage and 'Big Data' science (supporting a push for open government) in the United Kingdom and Canada, long term funding of data linkage infrastructure in Australia is required to avoid losing the competitive advantage that Australia has gained in the international data linkage arena and in the fields of research that use linked data.

## **Questions on resource costs of access**

*How should the costs associated with making more public sector data widely available be funded?*

Improved funding model. The funding environment is necessary to enable improvements to and expansion of services and delivery to a variety of user groups; to assist in prioritisation of activities. Without long term funding the infrastructure will be unable to realise its full potential. (Short-term planning/funding makes operation, maintenance and support of the infrastructure difficult and innovation virtually impossible).

*Is availability of skilled labour an issue in areas such as data science or other data-specific occupations? Is there a role for government in improving the skills base in this area?*

Expertise – data analytics requires expertise in three broad areas: knowledge of the datasets available for research (along with their characteristics and limitations), skills in data manipulation methods and skills in using/analysing information. By itself, a basic-level ability to use available data is insufficient, because correct interpretation of datasets depends on an understanding of the structure and content of, and variation within the component collections.

Programs like the NSW Biostatistician Training Program, established in 2000, provides broad training that enables graduates to apply biostatistical expertise to many different domains of public health practice. Graduates are skilled up to work as biostatisticians in a range of public health services, research, development, policy and planning positions.

## **Questions on privacy protection**

*What types of data and data applications (public sector and private sector) pose the greatest concerns for privacy protection?*

All university based research projects require ethical, custodian and institutional approval before they can proceed. At each stage, privacy and confidentiality restrictions add to the project-specific governance framework and control arrangements. The scope of the Information Governance Framework is to provide a systematic approach to safeguarding all sensitive information involved in the research activities. As a result, secure research facilities provide a safe environment to perform analysis on de-identified and/or appropriately confidentialised datasets.

*How can individuals' and businesses' confidence and trust in the way data is used be maintained and enhanced?*



Data custodians, researchers and record linkage centres have worked together to develop data access and usage models that comply with information privacy laws and provide necessary guards to privacy e.g. Australian Government High Level Principles for Data Integration (11). Moreover, record linkage units have implemented an array of best practice data governance policies to minimise the risk to privacy posed by their operations (1, 12-16).

Project-specific information governance encompasses people, processes, information technology (IT) systems, information and physical assets that support the research activities.

*What weight should be given to privacy protection relative to the benefits of greater data availability and use, particularly given the rate of change in the capabilities of technology?*

In an era of 'big data' development, there are significant challenges around data sharing and linkage. These include caution around data sharing and linkage and conservative interpretation of legislation around data release. There needs to be more thought given to the balance between data access, privacy and public benefit in research.

*Are further changes to the privacy-related policy framework needed? What are these specific changes and how would they improve outcomes? Have such approaches been tried in other jurisdictions?*

In most Western countries, information about an individual's health and welfare is collected as they come into contact with service delivery organisations and other government agencies, including hospitals (public and private), health departments and other human or social service authorities (e.g. education, criminal justice). Over time this data accumulates, providing a rich store of information that can be used to inform policy making and improve the health and social status of the entire community.

Technological advances have improved the accessibility, quality and integration potential of this data for research. In parallel, these 'big data' developments have helped establish flexible and transparent governance models that balance both privacy and the public interest in research. Developing proportionate governance frameworks based on clear guiding principles allows accurate assessment of risks associated with data use/sharing/linkage and assigns appropriate safeguards (17).

*How could coordination across the different jurisdictions in regard to privacy protection and legislation be improved?*

Many of the significant Commonwealth and State datasets are subject to specific legislation that defines the conditions of data release and/or use. The extent of this type of legislation and its complexity creates difficulties of interpretation with regard to the release of data for research projects. A truly consistent and transparent approach to data access and research assessment is required to ensure equity of data access.

*How effective are existing approaches to confidentialisation and data security in facilitating data sharing while protecting privacy?*

Existing approaches to confidentialisation and data security are often project specific and often restrictive. The governance of data for research needs to be simplified to allow agile responses to research and policy questions.

## Questions on data security

*Are security measures for public sector data too prescriptive? Do they need to be more flexible to adapt to changing circumstances and technologies?*

Security, ethics and privacy – in addition to legal requirements, access to many Commonwealth and State health datasets are subject to privacy and ethical review. The processes necessary to address privacy and confidentiality concerns are not always transparent. It should be clear what governance process, protocols and standards are required to enable safe and secure access to research data. In addition, the requirement for multiple ethics approvals (often in different application forms) adds additional layers of bureaucracy within the project approval process.

## Summary

Overall, data linkage infrastructure in Australia is recognised internationally for its high level of accuracy (linkage quality standards) and innovative technologies/methodologies. The challenge is to realise the potential of the infrastructure currently available across government and university sectors through compatible, sustainable and effective models which can maximise the capacity across all these systems.

Experiences from other countries demonstrate the need to harness and harmonise the power and experience of linkage services and systems to improve the efficiency and quality within overall linkage infrastructure.

## Attachment A - References

1. 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 services research*. 2012;12(1):480.
2. Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInnery K, Brown AP, et al. Accuracy and completeness of patient pathways—the benefits of national data linkage in Australia. *BMC Health Services Research*. 2015;15(1):312.
3. Spilsbury K, Rosman D, Alan J, Boyd J, Ferrante A, Semmens J. Cross border hospital use: analysis using data linkage across four Australian states. *The Medical journal of Australia*. 2015;202(11):582-6.
4. Diana Rosman, Katrina Spilsbury, Janine Alan, Anna Ferrante, Angela Young , Emma Fuller, et al. Multi-jurisdictional linkage in Australia: Proving a concept. *Australian and New Zealand Journal of Public Health*. 2014.
5. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. *Journal of Biomedical Informatics*. 2012;45(1):165-72.
6. Randall SM, Ferrante AM, Boyd JH, Semmens JB. The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making*. 2013;13(1):64.
7. Randall SM, Ferrante AM, Boyd JH, Semmens JB. Privacy-preserving record linkage on large real world datasets. *Journal of biomedical informatics*. 2013.
8. Randall SM, Boyd JH, Ferrante AM, Bauer JK, Semmens JB. Use of graph theory measures to identify errors in record linkage. *Computer methods and programs in biomedicine*. 2014;115(2):55-63.
9. Randall SM, Brown AP, Ferrante AM, Boyd JH, Semmens JB. Privacy preserving record linkage using homomorphic encryption. *Population Informatics for Big Data, Sydney, Australia*. 2015.
10. Boyd JH, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, et al. A Simple Sampling Method for Estimating the Accuracy of Large Scale Record Linkage Projects. *Methods Inf Med*. 2016;55(3):276-83.
11. Australian Government. High Level Principles for Data Integration involving Commonwealth Data for Statistical and Research Purposes. In: *Cross Portfolio Statistical Integration Committee (CPSIC), editor*. Canberra: Australian Government; 2010.
12. 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.
13. Harris J, editor *Next Generation Linkage Management System*. Sixth Australiasian Workshop on Health Informations and Knowledge Management; 2013; Adelaide, Australia: Australian Computer Society.
14. Trutwein B, Holman D, Rosman D. Health Data Linkage Conserves Privacy in a Research-Rich Environment. *Annals of Epidemiology*. 2006;16(4):279-80.
15. Ford D. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC health services research*. 2009;9(1):157.
16. Roos LL, Brownell M, Lix L, Roos NP, Walld R, L M. From health research to social research: Privacy, methods, approaches. *Social Science and Medicine*. 2008;66(1):117-29.
17. Laurie G, Sethi N. Towards principles-based approaches to governance of health-related research using personal data. *European journal of risk regulation: EJRR*. 2013;4(1):43.



## References

---

1. Dunn, H.L., *Record Linkage\**. American Journal of Public Health and the Nations Health, 1946. **36**(12): p. 1412-1416.
2. Hole, D., et al., *Cohort follow-up using computer linkage with routinely collected data*. Journal of chronic diseases, 1981. **34**(6): p. 291-297.
3. 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.
4. Clark, D., *Practical introduction to record linkage for injury research*. Injury Prevention, 2004. **10**(3): p. 186-191.
5. Moravec, H., *When will computer hardware match the human brain*. Journal of evolution and technology, 1998. **1**(1): p. 10.
6. Brown, B., M. Chui, and J. Manyika, *Are you ready for the era of 'big data'*. McKinsey Quarterly, 2011. **4**: p. 24-35.
7. Roos, L.L., et al., *From health research to social research: Privacy, methods, approaches*. Social science & medicine, 2008. **66**(1): p. 117-129.
8. Jorm, L., *Beyond Linkage: Using linked data for policy-relevant research*. Sydney. p. 20.
9. Jutte DP, Roos LL, and B. M, *Administrative record linkage as a tool for public health research*. Annual Review of Public Health 2011. **32**: p. 91-108.
10. Stanley F, et al., *Can Joined-up Data Lead to Joined-up Thinking? The Western Australian Developmental Pathways Project*. . Healthcare Policy, 2011. **6**: p. 68-79.
11. Berkeley, M., *TEXTBOOK OF MEDICAL RECORD LINKAGE*. The Journal of the Royal College of General Practitioners, 1987. **37**(304): p. 518.
12. Wajda, A., et al., *Record Linkage Strategies: Part II Portable Software and Deterministic Matching*. Methods of Information in Medicine, 1991. **30**: p. 210-214.
13. Gomatam, S., et al., *An Empirical Comparison of Record Linkage Procedures*. Statistics in Medicine, 2002. **21**: p. 1485-1496.
14. 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.
15. Newcombe, H.B., M.E. Fair, and P. Lalonde, *The Use of Names for Linking Personal Records*. Journal of the American Statistical Association, 1992. **87**: p. 1193-1204.
16. Newcombe, H.B., *Handbook for Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. 1988, New York: Oxford University Press. 210.
17. Newcombe, H. and J. Kennedy, *Record linkage: making maximum use of the discriminating power of identifying information*. . Commun. ACM, 1962. **5**(11): p. 563-566.
18. Basharin, G.P., *On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables*. Theory of Probability & Its Applications, 1959. **4**: p. 333-336.
19. Yancey, W.E., *Frequency-Dependent Probability Measures for Record Linkage*, U.B.o.t. Census, Editor. 2000: Washington DC. p. 6.
20. Roos, L.L., A. Wajda, and J.P. Nicol, *The art and science of record linkage: methods that work with few identifiers*. Computers and Biomedical Medicine, 1986. **16**(1): p. 45-47.
21. DuVall, S.L., R.A. Kerber, and A. Thomas, *Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators*. Journal of Biomedical Informatics, 2010. **43**: p. 24-30.
22. Gill, L.E., *OX-LINK: The Oxford Medical Record Linkage System*, in *Record Linkage Techniques*. 1997, University of Oxford: Oxford. p. 19.
23. Herzog, T.H., F. Scheuren, and W.E. Winkler, *Record Linkage*, in *Wires Computational Statistics*. 2010, John Wiley & Sons. p. 9.
24. 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.

25. Christen, P., *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Data-Centric Systems and Applications. 2012: Springer Science & Business Media.
26. Christen, P. and K. Goiser, *Quality and Complexity Measures for Data Linkage and Deduplication*, in *Quality Measures in Data Mining Studies in Computational Intelligence* F. Guillet and H. Hamilton, Editors. 2007, Springer. p. 127-151.
27. Clark, D.E. and D.R. Hahn. *Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry*. in *Proceedings from the Annual Symposium on Computer Application in Medical Care*. 1995.
28. Copas, J.B. and F.J. Hilton, *Record Linkage: Statistical Models for Matching Computer Records*. Journal of the Royal Statistical Society, 1990. **153**(3): p. 287-320.
29. Belin, T.R. and D.B. Rubin, *A Method for Calibrating False-Match Rates in Record Linkage*. Journal of the American Statistical Association, 2006. **90**(430): p. 694-707.
30. Bauman G John Jr, *Computation of Weights for Probabilistic Record Linkage using the EM Algorithm*. August 2006, Brigham Young University. p. 107.
31. Yancey, W., *Improving EM algorithm estimates for record linkage parameters*. Proceedings of the Section on Survey Research ..., 2002: p. 3835-3840.
32. Fellegi, I. and A. Sunter, *A Theory for Record Linkage*. Journal of the American Statistical Association, 1969. **64**: p. 1183-1210.
33. Winkler, W.E., *Advanced Methods for Record Linkage*, in *Statistical Research Report 1994*, U S Bureau of the Census, Statistical Research Division: Washington D C.
34. Winkler, W.E., *Approximate String Comparator Search Strategies for Very Large Administrative Lists*, U.S.B.o.t. Census, Editor. 2005: Washington, DC. p. 8.
35. 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. **84**(406): p. 414-420.
36. Chamberlayne, R., et al., *Creating a population-based linked health database: a new resource for health services research*. Canadian journal of public health. Revue canadienne de sante publique, 1997. **89**(4): p. 270-273.
37. Sibthorpe, B., E. Kliwer, and L. Smith, *Record linkage in Australian epidemiological research: Health benefits, privacy safeguards and future potential*. ANZ Journal of Public Health, 1995. **19**.
38. Roos, L.L. and A. Wajda, *Record Linkage Strategies: Part 1: Estimating Information and Evaluating Approaches*. 1990, University of Manitoba: Winnipeg. p. 28.
39. Kendrick, S.W. and J.A. Clarke, *The Scottish Medical Record Linkage System*. Health Bulletin (Edinburgh), 1979. **51**: p. 72-79.
40. Holman, D., 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**.
41. 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.
42. Ford, D.V., et al., *The SAIL Databank: building a national architecture for e-health research and evaluation* BMC Health Services Research 2009, 2009. **9**(157).
43. Acheson, E.D. and J.G. Evans, *The Oxford Record Linkage Study: A Review of the Method with some Preliminary Results*. Proc R Soc Med, 1964. **57**(4): p. 269-74.
44. Acheson, E.D., *Medical record linkage*. Medical record linkage., 1967.
45. Goldacre, M. *The value of linked data for policy development, strategic planning, clinical practice and public health: An international perspective*. in *Symposium on Health Data Linkage*. 2003. Public Health Information Development Unit, Adelaide University.

46. Gill, L. and M. Goldacre, *English national record linkage of hospital episode statistics and death registration records*. Unit of Health-Care Epidemiology, Oxford University, Oxford, 2003.
47. Martens, P.J. *Using the repository housed at the Manitoba centre for health policy: learning from the past, planning for the future*. in *Montreal, Quebec: Conference proceedings of the Statistics Canada Conference: Longitudinal Social and Health Surveys in an International Perspective*. 2006.
48. Hertzman, C.P., N. Meagher, and K.M. McGrail, *Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest*. *Journal of the American Medical Informatics Association*, 2013. **20**(1): p. 25-28.
49. Jones, K.H., et al., *A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: A privacy-protecting remote access system for health-related research and evaluation*. *Journal of Biomedical Informatics*, (0).
50. Hobbs, M. and M. McCall, *Health statistics and record linkage in Australia*. *Journal of Chronic Disease*, 1970. **23**: p. 375-381.
51. 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.
52. Kelman, C.W., A.J. Bass, and C.D.J. Holman, *Research use of linked health data - a best practice protocol*. *Australian and New Zealand Journal of Public Health*, 2002. **26**(3): p. 251-255.
53. 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.
54. Bouamrane, M.-M. and F. Mair. *An overview of electronic health systems development & integration in Scotland*. in *Proceedings of the first international workshop on Managing interoperability and complexity in health systems*. 2011. ACM.
55. Clark, D.N. *The Scottish Medical Record Linkage System: Past, Present and Future*. in *Exploiting Existing Data for Health Research*. 2009. St Andrews, Scotland.
56. Heasman, M., J. Donnelly, and V. Carstairs, *Analysis of data on practice of individual consultants*. *British journal of preventive & social medicine*, 1970. **24**(1): p. 64-65.
57. Heasman, M.A., *The Use of Record Linkage in Long-term Prospective Studies*. *Record Linkage in Medicine: Proceedings of the International Symposium Oxford, July 1967, 1968*.
58. Heasman, M.A. and J.A. Clarke, *Medical Record Linkage in Scotland*. *Health Bulletin (Edinburgh)*, 1979. **37**: p. 97-103.
59. Newcombe, H.B., et al., *Automatic Linkage of Vital Records*. *Science*, 1959: p. 954-959.
60. 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.
61. Kendrick, S. and J. Clarke, *The Scottish Record Linkage System*. *Health bulletin*, 1993. **51**(2): p. 72.
62. Kendrick, S. *The development of record linkage in Scotland: the responsive application of probability matching*. in *Proceedings of the 1997 record linkage workshop*. 1997.
63. Jaro, M.A., *Unimatch: A record linkage system: Users manual*. 1978: Bureau of the Census.
64. Kendrick, S.W. and R. McIlroy, *One Pass Linkage: The Rapid Creation of Patient-based Data*. *Proceedings of Healthcare Computing 1996, 1996*. **British Journal of Healthcare Computing Books: Weybridge, Surrey**.
65. Womersley, J., *The public health uses of the Scottish Community Health Index (CHI)*. *Journal of Public Health*, 1996. **18**(4): p. 465-472.
66. Mackay, D.F., et al., *Educational outcomes following breech delivery: a record-linkage study of 456 947 children*. *International Journal of Epidemiology*, 2015: p. dyu270.

67. Fleming, M., B. Kirby, and K.I. Penny, *Record linkage in Scotland and its applications to health research*. Journal of clinical nursing, 2012. **21**(19pt20): p. 2711-2721.
68. Trutwein, B., D. Holman, and D. Rosman, *Health Data Linkage Conserves Privacy in a Research-Rich Environment*. Annals of Epidemiology, 2006. **16**.
69. Frommer, P.M., et al., *NCRIS Capability 5.7: Population Health and Data Linkage 2007*, University of Sydney: Sydney. p. 8.
70. Boyd, J., et al., *Technical challenges of providing record linkage services for research*. BMC Medical Informatics and Decision Making, 2014. **14**(23).
71. Good, N., D. Hansen, and C.M. O'Keefe, *Linking and Analysing Health Data with Appropriate Privacy and Security*, in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. 2007: Amsterdam. p. 2.
72. Centre for Data Linkage, *Information Governance Framework for Data Linkage and Integration at the Centre for Data Linkage*. 2014, Curtin University.
73. Office of the Government Chief Information Officer, *Digital WA: Western Australian Government Information and Communications Technology (ICT) Strategy 2016 - 2020*, O.o.t.G.C.I. Officer, Editor. 2016, WA Government: Perth, Western Australia.
74. Australian Government, *Information and Communications Technology Security Manual (ism)*. 2006: Canberra. p. 257.
75. Adams, C. and J. Allen, *Government databases and public health research: Facilitating access in the public interest*. J Law Med, 2014. **21**(4): p. 957-72.
76. Allen J, et al., *Privacy protectionism and health information: is there any redress for harms to health?* Journal of Law and Medicine, 2013. **21**(2): p. 473-485.
77. Lyons, R.A., et al., *Development and use of a privacy-protecting total population record linkage system to support observational, interventional, and policy relevant research*. The Lancet, 2012. **380**, **Supplement 3**(0): p. S6.
78. Arai, M. and H. Tanaka, *A Proposal for an Effective Information Flow Control Model for Sharing and Protecting Sensitive Information*, in *Australasian Information Security Conference*. 2009, Australian Computer Society Inc: Wellington, New Zeland. p. 10.
79. Australian Government, *High Level Principles for Data Integration involving Commonwealth Data for Statistical and Research Purposes*, Cross Portfolio Statistical Integration Committee (CPSIC), Editor. 2010, Australian Government: Canberra.
80. Boyd, J., et al., *Data linkage infrastructure for cross-Jurisdictional health-related research in Australia*. BMC Health Services Research, 2012. **12**: p. 480.
81. Harris J. *Next Generation Linkage Management System*. in *Sixth Australasian Workshop on Health Informations and Knowledge Management*. 2013. Adelaide, Australia: Australian Computer Society.
82. Trutwein, B., D. Holman, and D. Rosman, *Health Data Linkage Conserves Privacy in a Research-Rich Environment*. Annals of Epidemiology, 2006. **16**(4): p. 279-280.
83. Ford, D., *The SAIL Databank: building a national architecture for e-health research and evaluation*. BMC health services research, 2009. **9**(1): p. 157.
84. Roos LL, et al., *From health research to social research: Privacy, methods, approaches*. Social Science and Medicine, 2008. **66**(1): p. 117-129.
85. Boyd JH, et al., *Data linkage infrastructure for cross-jurisdictional health-related research in Australia*. BMC Health Services Research, 2012. **12**.
86. Christen, P., *Privacy-Preserving Data Linkage and Geocoding: Current Approaches and Research Directions (ABSTRACT)*, in *Sixth IEEE International Conference on Data Mining - Workshops 2006*, IEEE Computer Society. p. 11.
87. Vatsalan, D., P. Christen, and V.S. Verykios, *A taxonomy of privacy-preserving record linkage techniques*. Information Systems, 2013. **38**(6): p. 946-969.
88. Vatsalan, D., P. Christen, and V.S. Verykios, *A taxonomy of privacy-preserving record linkage techniques*. Information Systems, 2013. **38**: p. 946-969.



89. Karakasidis, A. and V.S. Verykios, *Privacy preserving record linkage using phonetic codes*, in *2009 4th Balkan Conference in Informatics, BCI 2009*. 2009. p. 101-106.
90. Alhaqbani, B. and C. Fidge, *Privacy-preserving electronic health record linkage using pseudonym identifiers*, in *2008 10th IEEE Intl. Conf. on e-Health Networking, Applications and Service, HEALTHCOM 2008*. 2008. p. 108-117.
91. Bonomi, L., et al., *Privacy Preserving Record Linkage via grams Projections*. arXiv preprint arXiv:1208.2773, 2012.
92. Verykios, V.S. and P. Christen, *Privacy-preserving record linkage*, in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2013. p. 321-332.
93. Bachteler, T., J. Reiher, and R. Schnell, *Similarity Filtering with Multibit Trees for Record Linkage*. 2013.
94. Adrian Brown, C.B., Sean Randall and Rainer Schnell *High quality linkage using Multibit Trees for privacy-preserving blocking* in *2016 International Population Data Linkage Conference*. 2016, Swansea University: Swansea, Wales.
95. Schnell, R., T. Bachteler, and J. Reiher, *Privacy-preserving record linkage using Bloom filters*. *BMC Medical Informatics and Decision Making*, 2009. **9**(41).
96. Randall, S.M., et al., *Privacy-preserving record linkage on large real world datasets*. *Journal of biomedical informatics*, 2013.
97. 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.
98. Niedermeyer, F., et al., *Cryptanalysis of basic Bloom Filters used for Privacy Preserving Record Linkage*. 2014.
99. Schnell, R., et al., *Privacy-preserving record linkage using Bloom filters*. *BMC Medical Informatics and Decision Making*, 2009. **9**: p. 41.
100. 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.
101. Schnell, R., T. Bachteler, and J. Reiher, *A Novel Error-Tolerant Anonymous Linking Code*. 2011.
102. Schnell, R., *An efficient privacy-preserving record linkage technique for administrative data and censuses*. *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, 2014. **30**(3): p. 263-270.
103. Karapiperis, D. and V. Verykios, *An LSH-based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage*. *IEEE Transactions on Knowledge and Data Engineering*, 2014. **PP**: p. 1-1.
104. Lindell, Y. and B. Pinkas, *Secure multiparty computation for privacy-preserving data mining*. *Journal of Privacy and Confidentiality*, 2009. **1**(1): p. 5.
105. Vaikuntanathan, V., *Computing Blindfolded: New Developments in Fully Homomorphic Encryption*, in *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*. 2011, IEEE. p. 5-16.
106. Micciancio, D., *A first glimpse of cryptography's Holy Grail*. *Communications of the ACM*, 2010. **53**: p. 96.
107. Doröz, Y., E. Öztürk, and B. Sunar, *A million-bit multiplier architecture for fully homomorphic encryption*. *Microprocessors and Microsystems*, 2014. **38**(8, Part A): p. 766-775.
108. Vatsalan, D., et al., *An Evaluation Framework for Privacy-Preserving Record Linkage*. *Journal of Privacy and Confidentiality*, 2014. **6**.
109. 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.
110. Bachteler T, Schnell R, and Reiher J. *An Empirical comparison of approaches to approximate string matching in private record linkage*. in *Statistics Canada Symposium 2010. Social Statistics: The Interplay among Censuses, Surveys and Administrative Data*. 2010. Canada.

111. Kolb, L., A. Thor, and E. Rahm, *Load Balancing for MapReduce-based Entity Resolution*, in *2012 IEEE 28th International Conference on Data Engineering*. 2012, IEEE. p. 618-629.
112. Kolb, L., A. Thor, and E. Rahm, *Multi-pass sorted neighborhood blocking with MapReduce*. *Computer Science - Research and Development*, 2011. **27**: p. 45-63.
113. Kolb, L., A. Thor, and E. Rahm, *Block-based load balancing for entity resolution with MapReduce*, in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. 2011, ACM Press: New York, New York, USA. p. 2397.
114. Karapiperis, D. and V. Verykios, *A distributed near-optimal LSH-based framework for privacy-preserving record linkage*. *Computer Science and Information Systems*, 2014. **11**: p. 745-763.
115. Karapiperis, D., *A Distributed Framework For Scaling Up LSH-Based Computations in Privacy Preserving Record Linkage*, in *Proceedings of the 6th Balkan Conference in Informatics*. 2013. p. 102-109.
116. Guiver, T., *Sampling-Based Clerical Review Methods in Probabilistic Linking*. 2011, Australian Bureau of Statistics: ABS Website (<http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1351.0.55.034May%202011?OpenDocument>). p. 22.
117. Durham, E.A., et al., *Composite Bloom Filters for Secure Record Linkage*. *IEEE Transactions on Knowledge and Data Engineering*, 2014. **26**: p. 2956-2968.
118. Connell, F.A., P. Diehr, and L.G. Hart, *The use of large data bases in health care studies*. *Annual review of public health*, 1987. **8**(1): p. 51-74.
119. Harron, K., et al., *Evaluating bias due to data linkage error in electronic healthcare records*. *BMC medical research methodology*, 2014. **14**(1): p. 36.
120. Wright, J., G. Bishop, and T. Ayre, *Assessing the Quality of Linking Migrant Settlement Records to Census Data*, A.S. Branch, Editor. 2009, Australian Bureau of Statistics: Canberra. p. 44.
121. Bishop, G., *Assessing the Likely Quality of the Statistical Longitudinal Census Dataset*, A.B.o.S. Analytical Services Branch, Editor., Australian Bureau of Statistics: Canberra. p. 42.
122. Bishop, G. and J. Khoo, *Methodology of Evaluating the Quality of Probabilistic Linking*. 2007, Australian Bureau of Statistics, Analytical Services Branch: Canberra. p. 20.
123. Randall, S.M., et al., *Grouping methods for ongoing record linkage*, in *First International Workshop on Population Informatics for Big Data (PopInfo'15)*. 2015: Sydney, Australia.
124. Bishop, G., *Determining the quality of longitudinally linked Census data (Audiovisual material)*. 2008, Australian Bureau of Statistics: Canberra. p. 19.
125. Harron, K., et al., *Opening the black box of record linkage*. *Journal of epidemiology and community health*, 2012. **66**(12): p. 1198-1198.
126. Neter, J., E.S. Maynes, and R. Ramanathan, *The effect of mismatching on the measurement of response errors*. *Journal of the American Statistical Association*, 1965. **60**(312): p. 1005-1027.
127. 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.
128. Blakely, T. and C. Salmond, *Probabilistic record linkage and a method to calculate the positive predictive value*. *International Journal of Epidemiology*, 2002. **31**: p. 1246-1252.
129. Australian Bureau of Statistics, *Sampling Based Clerical Review Methods in Probabilistic Matching*, in *Statistics for informed decision making*.
130. Newcombe, H.B., et al., *Reliability of Computerized versus Manual Death Searches in a Study of the Health of Eldorado Uranium Workers*. *Computers in Biology and Medicine*, 1983. **13**(3): p. 157-169.
131. Bass, A.J. and C. Garfield, *Statistical linkage keys: How effective are they?*, in *Symposium on Health Data Linkage*. 2002, Available online at: <http://www.publichealth.gov.au/symposium.html>: Sydney 2002. p. 40-45.

132. AIHW, *The use of linkage keys for statistical work in community services: Background paper for the Statistical Linkage Project of the National Community Services Information Management Group*, A.I.o.H.a. Welfare, Editor. 2000: Canberra Australia.
133. Acemoglu, D. and A. Ozdaglar, *Graph Theory and Social Networks*. 2009: Massachusetts.
134. CHeReL, *Quality Assurance in Record Linkage*. 2009, Centre for Health Record Linkage (CHeReL): Sydney. p. 6.
135. Richards, D., V. Chellen, and P. Compton. *The reuse of ripple down rule knowledge bases: Using machine learning to remove repetition*. in *Proceedings of the 2nd Pacific Knowledge Acquisition Workshop (PKAW'96)*, Coogee, Australia. 1996. Citeseer.
136. Copeland, K.T., et al., *Bias due to misclassification in the estimation of relative risk*. American Journal of Epidemiology, 1977. **105**(5): p. 488-495.
137. 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.
138. Chambers, R., *Regression analysis of probability-linked data*. Official Statistics Research Series, 2009. **4**(2).
139. Kim, G. and R. Chambers, *Regression analysis under incomplete linkage*. Computational Statistics & Data Analysis, 2012. **56**(9): p. 2756-2770.
140. Kim, G. and R. Chambers, *Regression Analysis under Probabilistic Multi-Linkage*. Statistica Neerlandica, 2012. **66**(1): p. 64-79.
141. Lahiri, P. and M.D. Larsen, *Regression analysis with linked data*. Journal of the American statistical association, 2005. **100**(469): p. 222-230.
142. 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.
143. Harron, K., H. Goldstein, and C. Dibben, *Methodological Developments in Data Linkage*. 2015: John Wiley & Sons.
144. 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.
145. Spilsbury, K., et al., *Cross border hospital use: analysis using data linkage across four Australian states*. The Medical journal of Australia, 2015. **202**(11): p. 582-586.
146. Mitchell, R.J., et al., *Data linkage capabilities in Australia: practical issues identified by a Population Health Research Network 'Proof of Concept project'*. Australian and New Zealand Journal of Public Health, 2015: p. n/a-n/a.
147. Diana Rosman, et al., *Multi-jurisdictional linkage in Australia: Proving a concept*. Australian and New Zealand Journal of Public Health, 2014.
148. Winkler, W.E., *Matching and Record Linkage*, U.B.o.t. Census, Editor. 2005: Washington DC. p. 38.
149. Simmhan, Y., et al. *GrayWulf: Scalable Software Architecture for Data Intensive Computing*. in *42nd Hawaii International Conference on System Sciences*. 2009. Waikoloa, Big Island, Hawaii.
150. Flowers, J. and B. Ferguson, *The future of health intelligence: Challenges and opportunities*. Public health, 2010. **124**(5): p. 274-277.
151. Crilly, J.L., et al., *Linking ambulance, emergency department and hospital admissions data: understanding the emergency journey*. Medical Journal of Australia, 2011. **194**(4): p. S34-S37.
152. Ford, I. and W.o.S.C.P.S. Group, *Computerised Record Linkage: Compared with Traditional Patients follow-Up Methods in Clinical Trials and Illustrated in a Prospective Epidemiological Study*. Journal of Clinical Epidemiology, 1995. **48**(12): p. 1441-1452.
153. Semmens, J., et al., *The Quality of Surgical Care Project: A Model to Evaluate Surgical Outcomes in Western Australia Using Population-Based Record Linkage*. Australian and New Zealand Journal of Surgery, 1998. **68**(6): p. 397-403.

154. Hall, S.E., et al., *Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records*. International Journal for Quality in Health Care, 2005. **17**(5): p. 415-420.
155. Jarman, B., et al., *Monitoring changes in hospital standardised mortality ratios*. Bmj, 2005. **330**(7487): p. 329.
156. Dillner, L., *Scottish death rates published with health warning*. BMJ, 1994. **309**(6969): p. 1599-1600.
157. Duke, J., et al., *Burn Injury and cancer risk: A state-wide longitudinal study*. Burns, 2011. **38**: p. 340-47.
158. Marshall, M.N., et al., *The public release of performance data: what do we expect to gain? A review of the evidence*. Jama, 2000. **283**(14): p. 1866-1874.
159. Marshall, M.N., et al., *Public reporting on quality in the United States and the United Kingdom*. Health Affairs, 2003. **22**(3): p. 134-148.
160. Marshall, M., et al., *Public reporting of performance: lessons from the USA*. Journal of health services research & policy, 2000. **5**(1): p. 1.
161. MacAllan, L. and V. Narayan, *Keeping the♥ Beat in Grampian—a case study in community participation and ownership*. Health Promotion International, 1994. **9**(1): p. 13-19.
162. Hanlon, P., et al., *Hospital use by an ageing cohort: an investigation into the association between biological, behavioural and social risk markers and subsequent hospital utilization*. Journal of Public Health, 1998. **20**(4): p. 467-476.
163. Bradley, C.J., et al., *Health services research and data linkages: issues, methods, and directions for the future*. Health services research, 2010. **45**(5p2): p. 1468-1488.
164. 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.
165. Mitchell RJ, C. CM, and R. MR, *Data linkage for injury surveillance and research in Australia: perils, pitfalls and potential*. Australian and New Zealand Journal of Public Health, 2014. **38**(3): p. 275-280.
166. Pow, C., et al., *Privacy-Preserving Record Linkage: An international collaboration between Canada, Australia and Wales*, in *2016 International Population Data Linkage Conference*. 2016, Swansea University: Swansea, Wales.

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 1

---

### Conference Presentation Abstracts



## International Health Data Linkage Network (IHDLN) Conference

Perth, Australia, May 2012

### Development of the National Linkage System: a linkage system for the 21st Century

Boyd JH<sup>1\*</sup>, Ferrante AM<sup>1</sup>, Randall SM<sup>1</sup>, Bray J<sup>1</sup>, O'Shea M<sup>1</sup>, Semmens JB<sup>1</sup>

<sup>1</sup>Curtin University, Perth, AUSTRALIA

\* Presenter

#### Abstract

The Centre for Data Linkage (CDL) at Curtin University exists to conduct linkage of state and national datasets to service researchers. In order to do so effectively, an end-to-end production system is required to manage these linkages and extractions. While numerous software packages exist to handle ad hoc linkage of two files, there are few tools available to handle the more complex processes found in a production linkage system.

The National Linkage System (NLS) has been built to meet the needs of large scale routine matching for population health research. The system processes demographic information sent securely from data providers and adds it to the master file without manual input or intervention. Encrypted 'linkage system keys' are provided to data providers by request for approved research projects.

The system is designed to handle large volumes in short time frames. Additional information on the underlying relationships between recorded events is stored to enable sophisticated automated quality checks and fixes. The system is built to receive new, amended and deleted records, as well as to hold the history of all the relationships formed.

This talk will focus on some of the key features and design criteria for building an end to end production linkage system capable of handling large volumes.

# Scottish Informatics Programme (SHIP) International Conference: Exploiting existing data for health research

St Andrews, Scotland, September 2013

## Developing National Data Linkage Infrastructure in Australia

Boyd JH<sup>1\*</sup>, Ferrante AM<sup>1</sup>, Randall SM<sup>1</sup>, Bauer J<sup>1</sup>, Gillies M<sup>1</sup>, Semmens JB<sup>1</sup>

<sup>1</sup> Curtin University, Perth, AUSTRALIA

\* Presenter

### Abstract

The Centre for Data Linkage (CDL) at Curtin University in Perth, Western Australia was established to enable linkage of state and national datasets to service researchers. In order to do so effectively, the CDL has developed a series of compatible tools and modules in a linkage infrastructure framework.

Within this infrastructure framework, the CDL has developed an end-to-end production system to manage linkages and extractions, which handles the more complex processes typically found in a large scale production linkage unit. The CDL have also implemented a secure file transfer system to enable data delivery to and from this service. Finally, CDL have identified, and implemented a secure infrastructure system, designed to host this software, while ensuring security, privacy and the meeting of all legislative requirements.

This talk will focus on the process of designing and implementing this infrastructure, along with some of the key features found in each system.



# International Health Data Linkage Network (IHDLN) Conference Vancouver, Canada, April 2014

## Privacy preserving record linkage on large real world datasets

Boyd JH<sup>1\*</sup>, Randall SM<sup>1</sup>, Ferrante AM<sup>1</sup>, Bauer J<sup>1</sup>, Gillies M<sup>1</sup>, Semmens JB<sup>1</sup>

<sup>1</sup>Curtin University, Perth, AUSTRALIA

\* Presenter

### Abstract

**Objectives:** In this study we trial a method of record linkage using encrypted data which does not require the full disclosure of personally identifiable information.

**Approach:** The method uses encrypted personal identifying information (Bloom filters) in a probability-based linkage framework. The privacy preserving linkage method was tested using ten years of New South Wales (NSW) and Western Australian (WA) hospital admissions data, comprising in total over 26 million records.

**Results:** The linkage quality of the encrypted matching outputs was compared against results obtained using traditional probabilistic matching with full unencrypted personally identifying information. In our experimental setting, the results demonstrate that it is possible to link large volumes of data and achieve high quality linkage using encrypted data. Dataset 1, n= 6,772,949, Precision: Unencrypted Link 0.999, Encrypted Link 0.998, Recall: Unencrypted Link 0.981, Encrypted Link 0.981; Dataset 2, n= 19,874,083, Precision: Unencrypted Link 0.986, Encrypted Link 0.985, Recall: Unencrypted Link 0.972, Encrypted Link 0.970. While these results are encouraging, there are some issues which may be overcome for wider application.

**Conclusion:** This method shows promise and through adaptations of this method or similar privacy preserving methods, restrictions and risks related to information disclosure may be reduced so that the benefits of linked research can be fully realised.

# International Population Data Linkage Network (IPDLN) Conference Swansea, Wales, August 2016

## How do you measure up? Methods to assess linkage quality

Ferrante AM<sup>1</sup>, **Boyd JH**<sup>1\*</sup>, Randall SM<sup>1</sup>, Brown AP<sup>1</sup>, Semmens JB<sup>1</sup>

<sup>1</sup>Curtin University, Perth, AUSTRALIA

\* Presenter

### Abstract

**Objectives:** Record linkage is a powerful technique which transforms discrete episode data into longitudinal person-based records. These records enable the construction and analysis of complex pathways of health and disease progression, and service use. Achieving high linkage quality is essential for ensuring the quality and integrity of research based on linked data. The methods used to assess linkage quality will depend on the volume and characteristics of the datasets involved, the processes used for linkage and the additional information available for quality assessment. This paper proposes and evaluates two methods to routinely assess linkage quality.

**Approach:** Linkage units currently use a range of methods to measure, monitor and improve linkage quality; however, no common approach or standards exist. There is an urgent need to develop “best practices” in evaluating, reporting and benchmarking linkage quality. In assessing linkage quality, of primary interest is in knowing the number of true matches and non-matches identified as links and non-links. Any misclassification of matches within these groups introduces linkage errors. We present efforts to develop sharable methods to measure linkage quality in Australia. This includes a sampling-based method to estimate both precision (accuracy) and recall (sensitivity) following record linkage and a benchmarking method - a transparent and transportable methodology to benchmark the quality of linkages across different operational environments.

**Results:** The sampling-based method achieved estimates of linkage quality that were very close to actual linkage quality metrics. This method presents as a feasible means of accurately estimating matching quality and refining linkages in population level linkage studies. The benchmarking method provides a systematic approach to estimating linkage quality with a set of open and shareable datasets and a set of well-defined, established performance metrics. The method provides an opportunity to benchmark the linkage quality of different record linkage operations. Both methods have the potential to assess the inter-rater reliability of clerical reviews.

**Conclusions:** Both methods produce reliable estimates of linkage quality enabling the exchange of information within and between linkage communities. It is important that researchers can assess risk in studies using record linkage techniques. Understanding the impact of linkage quality on research outputs highlights a need for standard methods to routinely measure linkage quality. These two methods provide a good start to the quality process, but it is important to identify standards and good practices in all parts of the linkage process (pre-processing, standardising activities, linkage, grouping and extracting).

Conference proceeding published by the International Journal of Population Data Science

## International Population Data Linkage Network (IPDLN) Conference Swansea, Wales, August 2016

### Assessing the impact of different grouping methods: time to rethink and regroup?

Randall SM<sup>1</sup>, Ferrante AM<sup>1</sup>, Brown AP<sup>1</sup>, **Boyd JH<sup>1\*</sup>**, Semmens JB<sup>1</sup>

<sup>1</sup>Curtin University, Perth, AUSTRALIA

\* Presenter

#### Abstract

**Objectives:** The grouping of record-pairs to determine which administrative records belong to the same individual is an important process in record linkage. A variety of grouping methods are used but the relative benefits of each are unknown. We evaluate a number of grouping methods against the traditional merge based clustering approach using large scale administrative data.

**Approach:** The research aimed to both describe current grouping techniques used for record linkage, and to evaluate the most appropriate grouping method for specific circumstances. A range of grouping strategies were applied to three datasets with known truth sets. Conditions were simulated to appropriately investigate one-to-one, many-to-one and ongoing linkage scenarios.

**Findings:** Results suggest alternate grouping methods will yield large benefits in linkage quality, especially when the quality of the underlying repository is high. Stepwise grouping methods were clearly superior for one-to-one linkage. There appeared little difference in linkage quality between many-to-one grouping approaches. The most appropriate techniques for ongoing linkage depended on the quality of the population spine and the underlying dataset.

**Conclusions:** These results demonstrate the large effect that the choice of grouping strategy can have on overall linkage quality. Ongoing linkages to high quality population spines provide large improvements in linkage quality compared to merge based linkages. Procuring or developing such a population spine will provide high linkage quality at far lower cost than current methods for improving linkage quality. By improving linkage quality at low cost, this resource can be further utilised by health researchers.

Conference proceeding published by the International Journal of Population Data Science

# International Population Data Linkage Network (IPDLN) Conference Swansea, Wales, August 2016

## Using record linkage to examine long-term effects of burn injury: The Western Australian Population-based Burn Injury Project

Duke JM<sup>2</sup>, Boyd JH<sup>1\*</sup>, Randall SM<sup>1</sup>, Fear, MW<sup>2</sup>, Wood, FM<sup>2</sup>

<sup>1</sup>Curtin University, Perth, AUSTRALIA

<sup>2</sup>University of Western Australia, Perth, AUSTRALIA

<sup>3</sup>Fiona Wood Foundation, Perth, AUSTRALIA

\* Presenter

### Abstract

**Objectives:** While the most obvious impact of a burn is a visible scar, there are hidden impacts. The main contributors to adverse health outcomes after burns are the metabolic, inflammatory, immune and endocrine changes that occur in response to the initial injury. These responses have been shown to persist for at least three years after paediatric severe burns, with adverse effects to the circulatory and musculoskeletal systems. Recent evidence demonstrates that minor burns and severe burns can trigger these systemic responses. Currently, minimal data on the long-term effects of burns are available, and the data that do exist are primarily related to paediatric severe burns. We have used population-based record linkage to support a research program to shed light on the spectrum of long-term morbidity, expressed in terms of hospital admissions, experienced by burn patients to guide burn clinicians in the management of their patients. We report here our current findings of post-burn mortality and morbidity.

**Approach:** A population-based longitudinal study using linked hospital morbidity and death data from Western Australia was undertaken of all persons hospitalised for a first burn injury (n=30,997) in 1980-2012 and a frequency matched non-injury comparison cohort, randomly selected from Western Australia's birth registrations and electoral roll (n = 127,000). Crude admission rates and cumulative length of stay for disease-specific admissions were calculated. Negative binomial and Cox proportional hazards regression modelling were used to generate incidence rate ratios (IRR) and hazard ratios (HR), respectively, adjusting for sociodemographic and health factors.

**Results:** For both paediatric and adult burn patients we identified increased long-term all-cause mortality (IRR, 95%CI: <15 years: 1.6, 1.3-2.0; 15-44 years: 1.8, 1.7-2.0; ≥ 45 years: 1.4, 1.3-1.5). Increased post-burn discharge health service use for cardiovascular diseases (IRR, 95%CI: <15 years: 1.3, 1.1-1.6; 15-44 years: 1.6, 1.4-1.7; ≥ 45 years: 1.5, 1.4-1.6) and

musculoskeletal conditions (IRR, 95%CI: <20 years: 1.9, 1.7-2.1; ≥ 20 years: 2.0, 1.9-2.1) were also found. Analyses found significantly elevated admission rates for minor and severe burns. Adjusted HRs identified time periods after discharge where burn patients experienced significantly elevated disease-specific incident admissions (results not provided).

**Conclusion:** Both minor and severe burns were associated with increased long-term cardiovascular and musculoskeletal morbidity and mortality. These results identify treatment needs for burn patients for a prolonged time after discharge. Further research that links primary care and pharmaceutical data is required to facilitate identification of at-risk patients and appropriate treatment pathways to reduce post-burn morbidity.

Conference proceeding published by the International Journal of Population Data Science

## Appendix 2

---

### Statements of contribution






**Published manuscript**

**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 services research* 12.1 (2012): 480.**

*Contribution:*

JHB developed the research design and evaluation methodology for the paper, performed the evaluations, reviewed the literature, prepared 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:  \_\_\_\_\_

Christine O'Keefe: \_\_\_\_\_

John Bass: \_\_\_\_\_

Sean Randall:  \_\_\_\_\_

James Semmens:  \_\_\_\_\_

**Published manuscript**

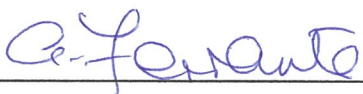
**Boyd JH, Randall SM, Ferrante AM, Bauer JK, Brown AP, Semmens JB. Technical challenges of providing record linkage services for research (2014). BMC Medical Informatics and Decision Making, 14 (1), art. no. 23.**

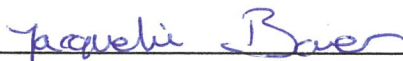
*Contribution:*

JHB developed the research design and model evaluation methodology for the paper, reviewed the literature, prepared the draft manuscript and edited the manuscript into its final form with the comments and suggestions of the other authors.

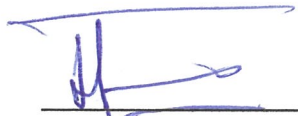
I acknowledge the above statement of contribution is accurate:

Sean Randall:  \_\_\_\_\_

Anna Ferrante:  \_\_\_\_\_

Jacqui Bauer:  \_\_\_\_\_

Adrian Brown:  \_\_\_\_\_

James Semmens:  \_\_\_\_\_

## Published manuscript

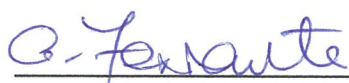
Ferrante AM and Boyd JH (2012). **A transparent and transportable methodology for evaluating Data Linkage software.** *Journal of Biomedical Informatics* (45)165-172.

### *Contribution:*

JHB was involved in developing the research design and evaluation methodology for the paper, performed evaluations, reviewed the literature, produced and interpreted results, supported preparation of the draft manuscript and provided comments and suggestion on the manuscript helping edited it into its final form.

I acknowledge the above statement of contribution is accurate:

Anna Ferrante:

  
\_\_\_\_\_

## Contributions to manuscripts – acknowledgements by co-authors

### Book Chapter


Boyd JH, Randall SM, Ferrante AM. **Application of privacy preserving techniques in operational record linkage centres.** *Medical Data Privacy Handbook. Springer International Publishing, 2015. 267-287.*

#### *Contribution:*

JHB reviewed the literature on operation linkage models, developed the research design and review methodology for the paper, prepared the first draft of the manuscript, edited the manuscript into its final form incorporating the comments and suggestions of the other authors.

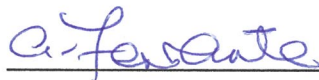
I acknowledge the above statement of contribution is accurate:

Sean Randall:



---

Anna Ferrante:



---

**Published manuscript**

Randall S, Ferrante AM, **Boyd JH**, Bauer JK, Semmens JB. **Privacy-preserving record linkage on large real world datasets.** *Journal of Biomedical Informatics* 2014; DOI: 10.1016/j.jbi.2013.12.003.

*Contribution:*

*JHB* supported development of the research design and evaluation methodology for the paper, and supported preparation of the draft manuscript providing comments and suggestion on the manuscript helping edited it into its final form.

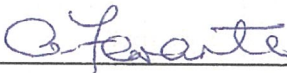
I acknowledge the above statement of contribution is accurate:

Sean Randall:



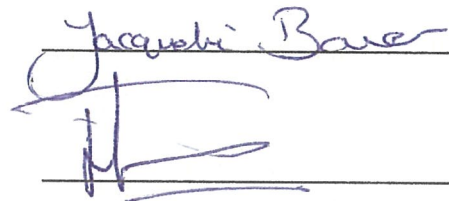
---

Anna Ferrante:



---

Jacqui Bauer:



---

James Semmens:



---


**Published manuscript**


**Boyd JH, Guiver T, Randall SM, Ferrante AM, Semmens JB, Anderson P, Dickinson T. A simple sampling method for estimating the accuracy of large scale record linkage projects. *Methods of information in medicine*. 2016;55(3):276-83.**

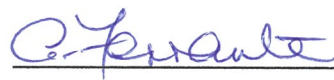
*Contribution:*

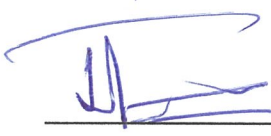
JHB developed the research design and evaluation methodology for the paper, performed the evaluations, reviewed the literature, produced and interpreted the results, prepared the first draft of the manuscript and edited the manuscript into its final form with the comments and suggestions of the other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

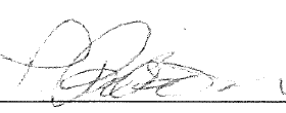
I acknowledge the above statement of contribution is accurate:

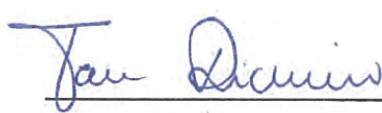
Tenniel Guiver: 

Sean Randall: 

Anna Ferrante: 

James Semmens: 

Phil Anderson: 

Teresa Dickinson:  7 NOV 2016

**Published manuscript**

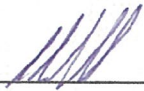
Randall S, Boyd JH, Ferrante AM, Semmens JB. **The effect of data cleaning on record linkage quality.** *BMC Medical Informatics and Decision Making* 2013; 13 (64): e1-e10.

*Contribution:*

**JHB** supported development of the research design and evaluation methodology to assess data cleaning measures, and supported preparation of the draft manuscript providing comments and suggestion on the manuscript helping edited it into its final form.

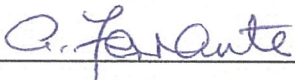
I acknowledge the above statement of contribution is accurate:

Sean Randall:



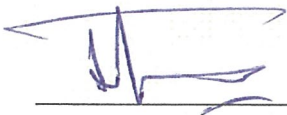
---

Anna Ferrante:



---

James Semmens:



---

**Published manuscript**

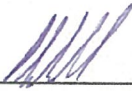
Randall S, Boyd JH, Ferrante AM, Semmens JB. **Use of graph theory measures to identify errors in record linkage.** *Computer Methods and Programs in Biomedicine*. Volume 115, Issue 2, July 2014, Pages 55-63.

*Contribution:*

*JHB* supported development of the research design and evaluation methodology for the graph theory measures, and supported preparation of the draft manuscript providing comments and suggestion on the manuscript helping edited it into its final form.

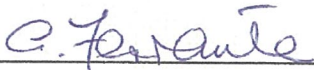
I acknowledge the above statement of contribution is accurate:

Sean Randall:




---

Anna Ferrante:



---

James Semmens:



---



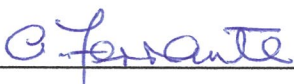
**Published letter**

**Boyd JH, Ferrante AM, Irvine K, Smith M, Moore E, Randall SM. Assessing linkage quality - what do researchers need to know? Australia and New Zealand Journal of Public Health.**

*Contribution:*

JHB developed the research design and evaluation methodology for the paper, reviewed the literature, prepared 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:  \_\_\_\_\_

Katie Irvine:  \_\_\_\_\_

Michael Smith:  \_\_\_\_\_

Elizabeth Moore:  \_\_\_\_\_

Sean Randall:  \_\_\_\_\_


**Published manuscript**

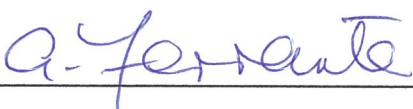
**Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInnery K, Brown AP, Spilsbury K, Gillies M and Semmens JB. Accuracy and completeness of patient pathways - the benefits of national data linkage in Australia. BMC health services research 15.1 (2015): 312.**

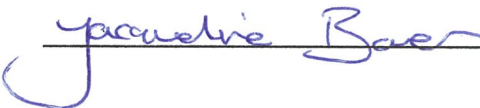
*Contribution:*

JHB developed the research design and evaluation methodology for the paper, performed the evaluations, reviewed the literature, produced and interpreted the results, prepared 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:


Sean Randall: 

Anna Ferrante: 

Jacqui Bauer: 

Kevin McInnery: 

Adrian Brown: 

Katrina Spilsbury: 

Margo Gillies: 

James Semmens: 

**Published manuscript**

Spilsbury K, Rosman D, Alan J, Ferrante AM, Boyd JH, Semmens JB. ***The Effect Of Cross-Jurisdictional Linked Administrative Data On Estimating Risk-Adjusted Grouped Hospital Standardised Mortality Ratios: A Retrospective Cohort Study.*** *Frontiers in Public Health* (2016).

*Contribution:*

*JHB* provided advice on linkage aspects of the research design, ethics and the draft manuscript; and had primary responsibility for development and execution of the linkage methodology for the study and interpretation of the linkage results for the manuscript.

I acknowledge the above statement of contribution is accurate:

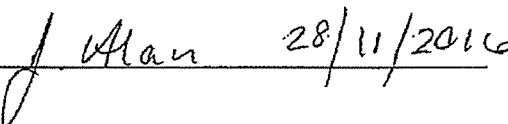
Katrina Spilsbury:

  
\_\_\_\_\_

Diana Rosman:

  
\_\_\_\_\_

Janine Alan:

  
\_\_\_\_\_

Anna Ferrante:

  
\_\_\_\_\_

James Semmens:

  
\_\_\_\_\_

**Published manuscript**

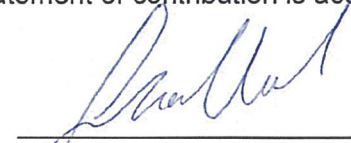
**Boyd JH, Wood FM, Randall SM, Fear MW, Rea S, Duke JM. (2016). Effects of pediatric burns on gastrointestinal diseases: A population-based study. *The Journal of Burn Care & Research*. July 2016.**

*Contribution:*

JHB was involved in developing the research design for the paper, reviewed the literature, presented and interpreted the results, prepared 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:

Fiona Wood:



---

Sean Randall:



---

Mark Fear:




---

Suzanne Rea:



---

Janine Duke:



---

**Published manuscript**

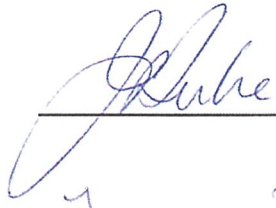
Duke JM, Bauer JK, Fear MW, Rea S, Wood FM, **Boyd JH.** (2014). **Burn injury, gender and cancer risk: population-based cohort study using data from Scotland and Western Australia.** *BMJ open*, 4(1), e003845.4

*Contribution:*

*JHB* was involved in developing the research design and evaluation methodology for the paper, performed evaluations, interpretation of results, supported preparation of the draft manuscript and provided comments and suggestion on the manuscript helping edited it into its final form.

I acknowledge the above statement of contribution is accurate:

Janine Duke:



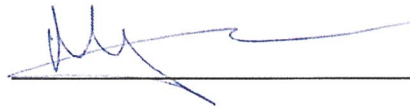
---

Jacqui Bauer:



---

Mark Fear:



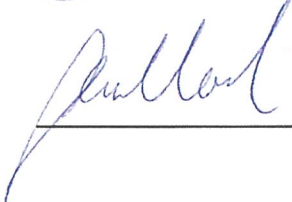
---

Suzanne Rea:



---

Fiona Wood:




---

# Statements of contribution

## *To Whom It May Concern*

I, James H Boyd, contributed to the advice and support, drafting of the article, reviewed and revised the manuscript, and approved the final manuscript as submitted

*Walsh D, Smalls M, and Boyd J. "Electronic health summaries-building on the foundation of Scottish Record Linkage system." Studies in health technology and informatics 2 (2001): 1212-1216.*



---

I, as the Lead Author, endorse that this level of contribution by the candidate indicated above is appropriate.



---

# Statements of contribution

## ***To Whom It May Concern***

I, James H Boyd, contributed to the advice and support, drafting of the articles, reviewed and revised the manuscripts, and approved the final manuscripts as submitted

*K MacIntyre, S Capewell, S Stewart, JWT Chalmers, J Boyd, A Finlayson, A Redpath, JP Pell and JJV McMurray. Evidence of improving prognosis in heart failure: trends in case fatality in 66 547 patients hospitalized between 1986 and 1995. Circulation 2000;102:1126-1131 doi: 10.1161/01.CIR.102.10.1126*

*K MacIntyre, S Stewart, S Capewell, JWT Chalmers, J Boyd, A Finlayson, A Redpath, H Gilmour, JJV McMurray. Gender and survival: a population-based study of 201,114 men and women following a first acute myocardial infarction. Journal of the American College of Cardiology Volume 38, Issue 3, September 2001; Pages 729-735 doi: 10.1016/S0735-1097(01)01465-6*

James Boyd

I, as the Lead Author, endorse that this level of contribution by the candidate indicated above is appropriate.

Kate MacIntyre KMacIntyre

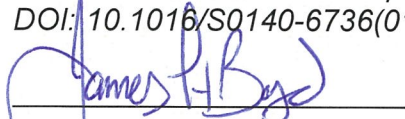
# Statements of contribution

## *To Whom It May Concern*

I, James H Boyd, contributed substantially to the advice and support, drafting of the articles, reviewed and revised the manuscripts, and approved the final manuscripts as submitted

*S Capewell, S Kendrick, J Boyd, G Cohen, E Juszczak, J Clarke. Measuring outcomes: one month survival after acute myocardial infarction in Scotland. Heart 1996; 76(1):70-5. DOI: 10.1016/S1062-1458(97)82162-9*

*S Capewell, K MacIntyre, S Stewart, JWT Chalmers, J Boyd, A Finlayson, A Redpath, J Pell, JJV McMurray. Age, sex, and social trends in out-of-hospital cardiac deaths in Scotland 1986-95: a retrospective cohort study. The Lancet 11/2001; 358(9289):1213-7. DOI: 10.1016/S0140-6736(01)06343-7*



---

I, as the Lead Author, endorse that this substantial level of contribution by the candidate indicated above is appropriate.



---

**Simon Capewell MD DSc**

*Professor of Clinical Epidemiology*

*University of Liverpool*

**Department of Public Health & Policy,**

Institute of Psychology, Health & Society.

Whelan Building, Quadrangle,

**University of Liverpool,**

LIVERPOOL, L69 3GB

United Kingdom

Telephone: 0044 (0)151 794 5576

Fax: 0044 (0)151 794 5588

Email: [capewell@liverpool.ac.uk](mailto:capewell@liverpool.ac.uk)



## Appendix 3

---

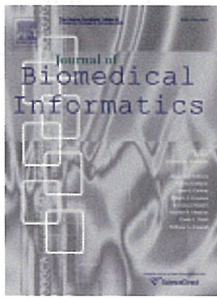
### Copyright statements

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.





# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


**Title:** A transparent and transportable methodology for evaluating Data Linkage software

**Author:** Anna Ferrante, James Boyd

**Publication:** Journal of Biomedical Informatics

**Publisher:** Elsevier

**Date:** February 2012

Copyright © 2011 Elsevier Inc. All rights reserved.

Logged in as:

James Boyd

Account #:  
3001084473

[LOGOUT](#)

## Order Completed

Thank you for your order.

This Agreement between James Boyd ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

### Printable details.

License Number	3991211297089
License date	Nov 17, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of Biomedical Informatics
Licensed Content Title	A transparent and transportable methodology for evaluating Data Linkage software
Licensed Content Author	Anna Ferrante, James Boyd
Licensed Content Date	February 2012
Licensed Content Volume	45
Licensed Content Issue	1
Licensed Content Pages	8
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	print
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research
Expected completion date	Jan 2017
Estimated size (number of pages)	300
Elsevier VAT number	GB 494 6272 12
Requestor Location	James Boyd Curtin University  Perth, 6102 Australia Attn: James Boyd
Total	0.00 AUD

[ORDER MORE](#)
[CLOSE WINDOW](#)



# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


**Title:** Application of Privacy-Preserving Techniques in Operational Record Linkage Centres

Logged in as:  
James Boyd

[LOGOUT](#)

**Author:** James H. Boyd

**Publication:** Springer eBook

**Publisher:** Springer

**Date:** Jan 1, 2015

Copyright © 2015, Springer International Publishing Switzerland

## Order Completed

Thank you for your order.

This Agreement between James Boyd ("You") and Springer ("Springer") consists of your license details and the terms and conditions provided by Springer and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

### [Printable details.](#)

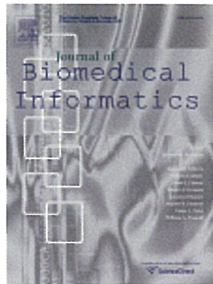
License Number	3991160047386
License date	Nov 17, 2016
Licensed Content Publisher	Springer
Licensed Content Publication	Springer eBook
Licensed Content Title	Application of Privacy-Preserving Techniques in Operational Record Linkage Centres
Licensed Content Author	James H. Boyd
Licensed Content Date	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	Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research
Expected completion date	Jan 2017
Estimated size(pages)	300
Requestor Location	James Boyd Curtin University  Perth, 6102 Australia Attn: James Boyd
Billing Type	Invoice
Billing address	James Boyd Curtin University  Perth, Australia 6102 Attn: James Boyd
Total	0.00 AUD

[ORDER MORE](#)
[CLOSE WINDOW](#)

Copyright © 2016 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement.](#) [Terms and Conditions.](#)  
Comments? We would like to hear from you. E-mail us at [customercare@copyright.com](mailto:customercare@copyright.com)



# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


**Title:** Privacy-preserving record linkage on large real world datasets

**Author:** Sean M. Randall, Anna M. Ferrante, James H. Boyd, Jacqueline K. Bauer, James B. Semmens

**Publication:** Journal of Biomedical Informatics

**Publisher:** Elsevier

**Date:** August 2014

Logged in as:  
James Boyd  
Account #:  
3001084473

[LOGOUT](#)

Copyright © 2013 Elsevier Inc. All rights reserved.

## Order Completed

Thank you for your order.

This Agreement between James Boyd ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

### Printable details.

License Number	3991220034919
License date	Nov 17, 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	50
Licensed Content Issue	n/a
Licensed Content Pages	8
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	print
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research
Expected completion date	Jan 2017
Estimated size (number of pages)	300
Elsevier VAT number	GB 494 6272 12
Requestor Location	James Boyd Curtin University  Perth, 6102 Australia Attn: James Boyd
Total	0.00 AUD



# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


**Title:** Use of graph theory measures to identify errors in record linkage

**Author:** Sean M. Randall, James H. Boyd, Anna M. Ferrante, Jacqueline K. Bauer, James B. Semmens

**Publication:** Computer Methods and Programs in Biomedicine

**Publisher:** Elsevier

**Date:** July 2014

Logged in as:  
James Boyd  
Account #:  
3001084473

[LOGOUT](#)

Copyright © 2014 Elsevier Ireland Ltd. All rights reserved.

## Order Completed

Thank you for your order.

This Agreement between James Boyd ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

### Printable details.

License Number	3991220303343
License date	Nov 17, 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	115
Licensed Content Issue	2
Licensed Content Pages	9
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	print
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research
Expected completion date	Jan 2017
Estimated size (number of pages)	300
Elsevier VAT number	GB 494 6272 12
Requestor Location	James Boyd Curtin University  Perth, 6102 Australia Attn: James Boyd
Total	0.00 AUD

[ORDER MORE](#)
[CLOSE WINDOW](#)



# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


**Title:** Understanding the origins of record linkage errors and how they affect research outcomes

**Author:** James H. Boyd, Anna M. Ferrante, Katie Irvine, Michael Smith, Elizabeth Moore, Adrian Brown, Sean M. Randall

**Publication:** Australian and New Zealand Journal of Public Health

**Publisher:** John Wiley and Sons

**Date:** Nov 20, 2016

© 2016 Public Health Association of Australia

Logged in as:

James Boyd

 Account #:  
3001084473

[LOGOUT](#)

## Order Completed

Thank you for your order.

This Agreement between James Boyd ("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.

Your confirmation email will contain your order number for future reference.

### [Printable details.](#)

License Number	3993381415813
License date	Nov 20, 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
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research
Expected completion date	Jan 2017
Expected size (number of pages)	300
Requestor Location	James Boyd Curtin University
	Perth, 6102 Australia Attn: James Boyd
Publisher Tax ID	EU826007151
Billing Type	Invoice
Billing address	James Boyd Curtin University
	Perth, Australia 6102 Attn: James Boyd
Total	0.00 AUD

**Would you like to purchase the full text of this article? If so, please continue on to the**



RightsLink®

Home

Account  
Info

Help



**Title:** Effects of Pediatric Burns on  
Gastrointestinal Diseases: A  
Population-Based Study.

**Author:** James Boyd, Fiona Wood, Sean  
Randall, et al

**Publication:** Journal of Burn Care & Research

**Publisher:** Wolters Kluwer Health, Inc.

**Date:** Aug 16, 0929

Copyright © 2016, (C) 2016 The American Burn  
Association

Logged in as:  
James Boyd  
Account #:  
3001084473

LOGOUT

This reuse is free of charge. No permission letter is needed from Wolters Kluwer Health, Lippincott Williams & Wilkins. We require that all authors always include a full acknowledgement. Example: AIDS: 13 November 2013 - Volume 27 - Issue 17 - p 2679-2689. Wolters Kluwer Health Lippincott Williams & Wilkins© No modifications will be permitted.

BACK

CLOSE WINDOW

Copyright © 2016 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement.](#) [Terms and Conditions.](#)  
Comments? We would like to hear from you. E-mail us at [customercare@copyright.com](mailto:customercare@copyright.com)





# RightsLink®

[Home](#)[Account Info](#)[Help](#)**Wolters Kluwer****Title:** Evidence of Improving Prognosis in Heart Failure**Author:** K. MacIntyre,S. Capewell,S. Stewart,J.W.T. Chalmers,J. Boyd,A. Finlayson,A. Redpath,J.P. Pell,J.J.V. McMurray**Publication:** Circulation**Publisher:** Wolters Kluwer Health, Inc.**Date:** Sep 5, 2000

Copyright © 2000, American Heart Association, Inc.

Logged in as:

James Boyd

Account #:  
3001084473[LOGOUT](#)

This reuse is free of charge. No permission letter is needed from Wolters Kluwer Health, Lippincott Williams & Wilkins. We require that all authors always include a full acknowledgement. Example: AIDS: 13 November 2013 - Volume 27 - Issue 17 - p 2679-2689. Wolters Kluwer Health Lippincott Williams & Wilkins© No modifications will be permitted.

[BACK](#)[CLOSE WINDOW](#)

Copyright © 2016 [Copyright Clearance Center, Inc.](#) All Rights Reserved. [Privacy statement.](#) [Terms and Conditions.](#) Comments? We would like to hear from you. E-mail us at [customercare@copyright.com](mailto:customercare@copyright.com)



# RightsLink®

[Home](#)
[Account Info](#)
[Help](#)


## THE LANCET



**Title:** Age, sex, and social trends in out-of-hospital cardiac deaths in Scotland 1986–95: a retrospective cohort study

**Author:** Simon Capewell, Kate MacIntyre, Simon Stewart, Jim WT Chalmers, James Boyd, Alan Finlayson, Adam Redpath, Jill P Pell, John JV McMurray

**Publication:** The Lancet

**Publisher:** Elsevier

**Date:** 13 October 2001

Copyright © 2001 Elsevier Ltd. All rights reserved.

Logged in as:

James Boyd

Account #:

3001084473

[LOGOUT](#)

## Order Completed

Thank you for your order.

This Agreement between James Boyd ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

Your confirmation email will contain your order number for future reference.

### Printable details.

License Number	3991210996224
License date	Nov 17, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	The Lancet
Licensed Content Title	Age, sex, and social trends in out-of-hospital cardiac deaths in Scotland 1986–95: a retrospective cohort study
Licensed Content Author	Simon Capewell, Kate MacIntyre, Simon Stewart, Jim WT Chalmers, James Boyd, Alan Finlayson, Adam Redpath, Jill P Pell, John JV McMurray
Licensed Content Date	13 October 2001
Licensed Content Volume	358
Licensed Content Issue	9289
Licensed Content Pages	5
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	print
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research
Expected completion date	Jan 2017
Estimated size (number of pages)	300
Elsevier VAT number	GB 494 6272 12
Requestor Location	James Boyd Curtin University

Perth, 6102



RightsLink®

Account  
Info

Help



**Title:** Gender and survival: a population-based study of 201,114 men and women following a first acute myocardial infarction

**Author:** Kate MacIntyre, Simon Stewart, Simon Capewell, James W.T Chalmers, Jill P Pell, James Boyd, Alan Finlayson, Adam Redpath, Harper Gilmour, John J.V McMurray

**Publication:** Journal of the American College of Cardiology

**Publisher:** Elsevier

**Date:** Sep 1, 2001

Copyright © 2001, Elsevier

Logged in as:  
James Boyd  
Account #:  
3001084473

LOGOUT

## Order Completed

Thank you for your order.

This Agreement between James Boyd ("You") and Elsevier ("Elsevier") consists of your order details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License number	Reference confirmation email for license number
License date	Nov 17, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of the American College of Cardiology
Licensed Content Title	Gender and survival: a population-based study of 201,114 men and women following a first acute myocardial infarction
Licensed Content Author	Kate MacIntyre, Simon Stewart, Simon Capewell, James W.T Chalmers, Jill P Pell, James Boyd, Alan Finlayson, Adam Redpath, Harper Gilmour, John J.V McMurray
Licensed Content Date	September 2001
Licensed Content Volume	38
Licensed Content Issue	3
Licensed Content Pages	7
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	print
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research
Expected completion date	Jan 2017
Estimated size (number of pages)	300
Elsevier VAT number	GB 494 6272 12
Requestor Location	James Boyd Curtin University

Perth, 6102