Contents

Foreword xi
Contributors xiii

1 Introduction 1
Katie Harron, Harvey Goldstein and Chris Dibben

1.1 Introduction: data linkage as it exists 1
1.2 Background and issues 2
1.3 Data linkage methods 3
  1.3.1 Deterministic linkage 3
  1.3.2 Probabilistic linkage 3
  1.3.3 Data preparation 4
1.4 Linkage error 5
1.5 Impact of linkage error on analysis of linked data 6
1.6 Data linkage: the future 7

2 Probabilistic linkage 8
William E. Winkler

2.1 Introduction 8
2.2 Overview of methods 10
  2.2.1 The Fellegi–Sunter model of record linkage 10
  2.2.2 Learning parameters 13
  2.2.3 Additional methods for matching 20
  2.2.4 An empirical example 22
2.3 Data preparation 23
  2.3.1 Description of a matching project 24
  2.3.2 Initial file preparation 25
  2.3.3 Name standardisation and parsing 26
  2.3.4 Address standardisation and parsing 27
  2.3.5 Summarising comments on preprocessing 27
2.4 Advanced methods 28
  2.4.1 Estimating false-match rates without training data 28
  2.4.2 Adjusting analyses for linkage error 32
2.5 Concluding comments 35
3  The data linkage environment

*Chris Dibben, Mark Elliot, Heather Gowans, Darren Lightfoot and Data Linkage Centres*

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>36</td>
</tr>
<tr>
<td>3.2 The data linkage context</td>
<td>37</td>
</tr>
<tr>
<td>3.2.1 Administrative or routine data</td>
<td>37</td>
</tr>
<tr>
<td>3.2.2 The law and the use of administrative (personal) data for research</td>
<td>38</td>
</tr>
<tr>
<td>3.2.3 The identifiability problem in data linkage</td>
<td>42</td>
</tr>
<tr>
<td>3.3 The tools used in the production of functional anonymity through a data linkage environment</td>
<td>42</td>
</tr>
<tr>
<td>3.3.1 Governance, rules and the researcher</td>
<td>43</td>
</tr>
<tr>
<td>3.3.2 Application process, ethics scrutiny and peer review</td>
<td>43</td>
</tr>
<tr>
<td>3.3.3 Shaping ‘safe’ behaviour: training, sanctions, contracts and licences</td>
<td>43</td>
</tr>
<tr>
<td>3.3.4 ‘Safe’ data analysis environments</td>
<td>44</td>
</tr>
<tr>
<td>3.3.5 Fragmentation: separation of linkage process and temporary linked data</td>
<td>47</td>
</tr>
<tr>
<td>3.4 Models for data access and data linkage</td>
<td>50</td>
</tr>
<tr>
<td>3.4.1 Single centre</td>
<td>50</td>
</tr>
<tr>
<td>3.4.2 Separation of functions: firewalls within single centre</td>
<td>51</td>
</tr>
<tr>
<td>3.4.3 Separation of functions: TTP linkage</td>
<td>53</td>
</tr>
<tr>
<td>3.4.4 Secure multiparty computation</td>
<td>53</td>
</tr>
<tr>
<td>3.5 Four case study data linkage centres</td>
<td>54</td>
</tr>
<tr>
<td>3.5.1 Population Data BC</td>
<td>54</td>
</tr>
<tr>
<td>3.5.2 The Secure Anonymised Information Linkage Databank, United Kingdom</td>
<td>58</td>
</tr>
<tr>
<td>3.5.3 Centre for Data Linkage (Population Health Research Network), Australia</td>
<td>59</td>
</tr>
<tr>
<td>3.5.4 The Centre for Health Record Linkage, Australia</td>
<td>61</td>
</tr>
<tr>
<td>3.6 Conclusion</td>
<td>62</td>
</tr>
</tbody>
</table>

4  Bias in data linkage studies

*Megan Bohensky*

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Background</td>
<td>63</td>
</tr>
<tr>
<td>4.2 Description of types of linkage error</td>
<td>65</td>
</tr>
<tr>
<td>4.2.1 Missed matches from missing linkage variables</td>
<td>65</td>
</tr>
<tr>
<td>4.2.2 Missed matches from inconsistent case ascertainment</td>
<td>66</td>
</tr>
<tr>
<td>4.2.3 False matches: Description of cases incorrectly matched</td>
<td>66</td>
</tr>
<tr>
<td>4.3 How linkage error impacts research findings</td>
<td>68</td>
</tr>
<tr>
<td>4.3.1 Results</td>
<td>68</td>
</tr>
<tr>
<td>4.3.2 Assessment of linkage bias</td>
<td>75</td>
</tr>
<tr>
<td>4.4 Discussion</td>
<td>78</td>
</tr>
<tr>
<td>4.4.1 Potential biases in the review process</td>
<td>79</td>
</tr>
<tr>
<td>4.4.2 Recommendations and implications for practice</td>
<td>79</td>
</tr>
</tbody>
</table>
5 Secondary analysis of linked data
Raymond Chambers and Gunky Kim

5.1 Introduction
5.2 Measurement error issues arising from linkage
5.2.1 Correct links, incorrect links and non-links
5.2.2 Characterising linkage errors
5.2.3 Characterising errors from non-linkage
5.3 Models for different types of linking errors
5.3.1 Linkage errors under binary linking
5.3.2 Linkage errors under multi-linking
5.3.3 Incomplete linking
5.3.4 Modelling the linkage error
5.4 Regression analysis using complete binary-linked data
5.4.1 Linear regression
5.4.2 Logistic regression
5.5 Regression analysis using incomplete binary-linked data
5.5.1 Linear regression using incomplete sample to register linked data
5.6 Regression analysis with multi-linked data
5.6.1 Uncorrelated multi-linking: Complete linkage
5.6.2 Uncorrelated multi-linking: Sample to register linkage
5.6.3 Correlated multi-linkage
5.6.4 Incorporating auxiliary population information
5.7 Conclusion and discussion

6 Record linkage: A missing data problem
Harvey Goldstein and Katie Harron

6.1 Introduction
6.2 Probabilistic Record Linkage (PRL)
6.3 Multiple Imputation (MI)
6.4 Prior-Informed Imputation (PII)
6.4.1 Estimating matching probabilities
6.5 Example 1: Linking electronic healthcare data to estimate trends in bloodstream infection
6.5.1 Methods
6.5.2 Results
6.5.3 Conclusions
6.6 Example 2: Simulated data including non-random linkage error
6.6.1 Methods
6.6.2 Results
6.7 Discussion
6.7.1 Non-random linkage error
6.7.2 Strengths and limitations: Handling linkage error
6.7.3 Implications for data linkers and data users
7 Using graph databases to manage linked data 125
James M. Farrow

7.1 Summary 125

7.2 Introduction 126
7.2.1 Flat approach 127
7.2.2 Oops, your legacy is showing 128
7.2.3 Shortcomings 128

7.3 Graph approach 131
7.3.1 Overview of graph concepts 131
7.3.2 Graph queries versus relational queries 133
7.3.3 Comparison of data in flat database versus graph database 136
7.3.4 Relaxing the notion of ‘truth’ 137
7.3.5 Not a linkage approach per se but a management approach which enables novel linkage approaches 138
7.3.6 Linkage engine independent 139
7.3.7 Separates out linkage from cluster identification phase (and clerical review) 139

7.4 Methodologies 139
7.4.1 Overview of storage and extraction approach 140
7.4.2 Overall management of data as collections 141
7.4.3 Data loading 142
7.4.4 Identification of equivalence sets and deterministic linkage 143
7.4.5 Probabilistic linkage 144
7.4.6 Clerical review 144
7.4.7 Determining cut-off thresholds 145
7.4.8 Final cluster extraction 147
7.4.9 Graph partitioning 147
7.4.10 Data management/curation 150
7.4.11 User interface challenges 150
7.4.12 Final cluster extraction 154
7.4.13 A typical end-to-end workflow 155

7.5 Algorithm implementation 156
7.5.1 Graph traversal 156
7.5.2 Cluster identification 157
7.5.3 Partitioning visitor 158
7.5.4 Encapsulating edge following policies 158
7.5.5 Graph partitioning 158
7.5.6 Insertion of review links 158
7.5.7 How to migrate while preserving current clusters 158

7.6 New approaches facilitated by graph storage approach 158
7.6.1 Multiple threshold extraction 160
7.6.2 Possibility of returning graph to end users 165
7.6.3 Optimised cluster analysis 166
7.6.4 Other link types 167

7.7 Conclusion 167
8 Large-scale linkage for total populations in official statistics

Owen Abbott, Peter Jones and Martin Ralphs

8.1 Introduction
8.2 Current practice in record linkage for population censuses
  8.2.1 Introduction
  8.2.2 Case study: the 2011 England and Wales Census assessment of coverage
8.3 Population-level linkage in countries that operate a population register: register-based censuses
  8.3.1 Introduction
  8.3.2 Case study 1: Finland
  8.3.3 Case study 2: The Netherlands Virtual Census
  8.3.4 Case study 3: Poland
  8.3.5 Case study 4: Germany
  8.3.6 Summary
8.4 New challenges in record linkage: the Beyond 2011 Programme
  8.4.1 Introduction
  8.4.2 Beyond 2011 linking methodology
  8.4.3 The anonymisation process in Beyond 2011
  8.4.4 Beyond 2011 linkage strategy using pseudonymised data
  8.4.5 Linkage quality
  8.4.6 Next steps
  8.4.7 Conclusion
8.5 Summary

9 Privacy-preserving record linkage

Rainer Schnell

9.1 Introduction
9.2 Chapter outline
9.3 Linking with and without personal identification numbers
  9.3.1 Linking using a trusted third party
  9.3.2 Linking with encrypted PIDs
  9.3.3 Linking with encrypted quasi-identifiers
  9.3.4 PPRL in decentralised organisations
9.4 PPRL approaches
  9.4.1 Phonetic codes
  9.4.2 High-dimensional embeddings
  9.4.3 Reference tables
  9.4.4 Secure multiparty computations for PPRL
  9.4.5 Bloom filter-based PPRL
9.5 PPRL for very large databases: blocking
  9.5.1 Blocking for PPRL with Bloom filters
  9.5.2 Blocking Bloom filters with MBT
  9.5.3 Empirical comparison of blocking techniques for Bloom filters