Seeing the big picture - the importance of taking an 'end-to-end' view on administrative data linkage

Chris Dibben
University of Edinburgh

THE UNIVERSITY of EDINBURGH

Creating Research Data Policy in a Changing Data Landscape

Acknowledgements

- *Andy Boyd, Peter Christen, Chris Dibben, Mark Elliot, Christine M O'Keefe*
- Data linkage and anonymization – Isaac Newton Institute for Mathematical Sciences

Overview

- Motivation – administrative data linkage
- Traditional approaches
- Benefits of a more expansive/ flexible approach
- A more general way to think about this ‘end-to-end’
- A collision with privacy

83RD EDITION

WISDEN

CRICKETERS' ALMANACK

1946

EDITED BY HUBERT PRESTON

SPECIAL FEATURES

EDWARD PAYNTER
by R. C. Robertson-Glasgow

CRICKET UNDER THE JAPS
by E. W. Swanton

A HUNDRED YEARS OF SURREY CRICKET
by H. G. D. Leveson-Gower

NOTES BY THE EDITOR

PUBLISHED BY SPORTING HANDBOOKS LTD
AT 13 BEDFORD SQUARE LONDON WC1 FOR
THE PROPRIETORS JOHN WISDEN AND CO LTD

Administrative Data

- collected by government departments and other organisations…
  - registration, transaction and record keeping,
  - delivering a service or for day-to-day operations
  - not research-ready
- cover a population
- important new resource for social scientists
  - coverage, methodology
  - better understanding of our society
  - better informed government policy

People with autism 'die younger', warns charity

By Dominic Howell
BBC News

18 March 2016 | Health

Thinkstock

Top Stories

EU leaders put migrant deal to Turkey
EU leaders hold talks with Turkey's prime minister in an attempt to reach a deal over the migrant crisis.
1 hour ago

No 'concessions' over disability cuts
2 minutes ago

Ben Nevis gains a metre thanks to GPS
18 March 2016

Features

BJPsych The British Journal of Psychiatry (2016)
208, 232–238. doi: 10.1192/bjp.bp.114.160192

Premature mortality in autism spectrum disorder

Tatja Hirvikoski, Ellenor Mittendorfer-Rutz, Marcus Boman, Henrik Larsson, Paul Lichtenstein and Sven Bölte

Background
Mortality has been suggested to be increased in autism spectrum disorder (ASD).

Aims
To examine both all-cause and cause-specific mortality in ASD, as well as investigate moderating role of gender and intellectual ability.

Method
Odds ratios (ORs) were calculated for a population-based cohort of ASD probands ($n=27\,122$, diagnosed between 1987 and 2009) compared with gender-, age- and county of residence-matched controls ($n=2\,672\,185$).

Results
During the observed period, 24 358 (0.91%) individuals in the general population died, whereas the corresponding figure for individuals with ASD was 706 (2.60%; OR = 2.56; 95% CI 2.38–2.76). Cause-specific analyses showed elevated mortality in ASD for almost all analysed diagnostic categories. Mortality and patterns for cause-specific mortality were partly moderated by gender and general intellectual ability.

Conclusions
Premature mortality was markedly increased in ASD owing to a multitude of medical conditions.

Declaration of interest
None.

Table 3 Risk for all-cause mortality for the entire autism spectrum disorder (ASD) group, as well as separately for females and males, and low-functioning ASD and high-functioning ASD groups

	Controls Number of deaths (%)	ASD OR (95% CI) Number of deaths (%)	Low-functioning ASD OR (95% CI) Number of deaths (%)	High-functioning ASD OR (95% CI) Number of deaths (%)
Total	24 358 (0.91)	2.56 (2.38–2.76) 706 (2.60)	5.78** (4.94–6.75) 169 (2.71)	2.18 (2.00–2.38) 537 (2.57)
Females	11 693 (1.39)	2.24 (1.99–2.51) 296 (3.51)	8.52 (6.55–11.08) 61 (3.00)	1.88 (1.65–2.14) 235 (3.67)
Males	12 665 (0.69)	2.87* (2.60–3.16) 410 (2.19)	4.88 (4.02–5.93) 108 (2.57)	2.49 (2.22–2.80) 302 (2.08)

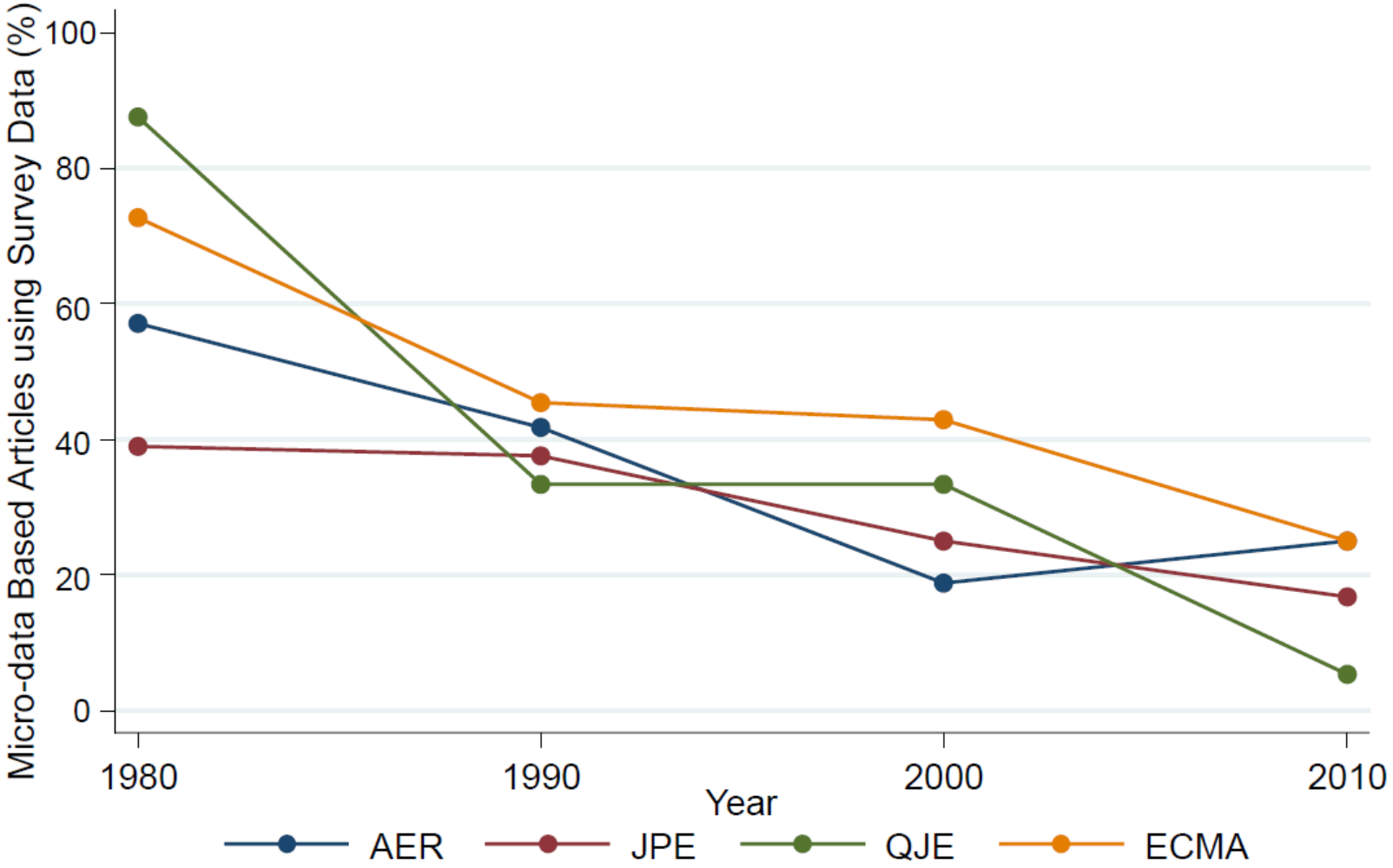
ASD, autism spectrum disorder; OR, odds ratio; CI, confidence interval.
*Partial likelihood ratio test for interaction effect ASD × gender, $P=0.001$.
**Partial likelihood ratio test for model selection (low-functioning ASD/high-functioning ASD), $P<0.001$.

Why use linked administrative data for research?

- Relatively cheap form of research
- Often cover 100% of a population – small groups in the population, rare events.
- Historical – can re-explore what happened

- Useful: events, variation, small area geographic patterns, natural experiments/ quasi experiments

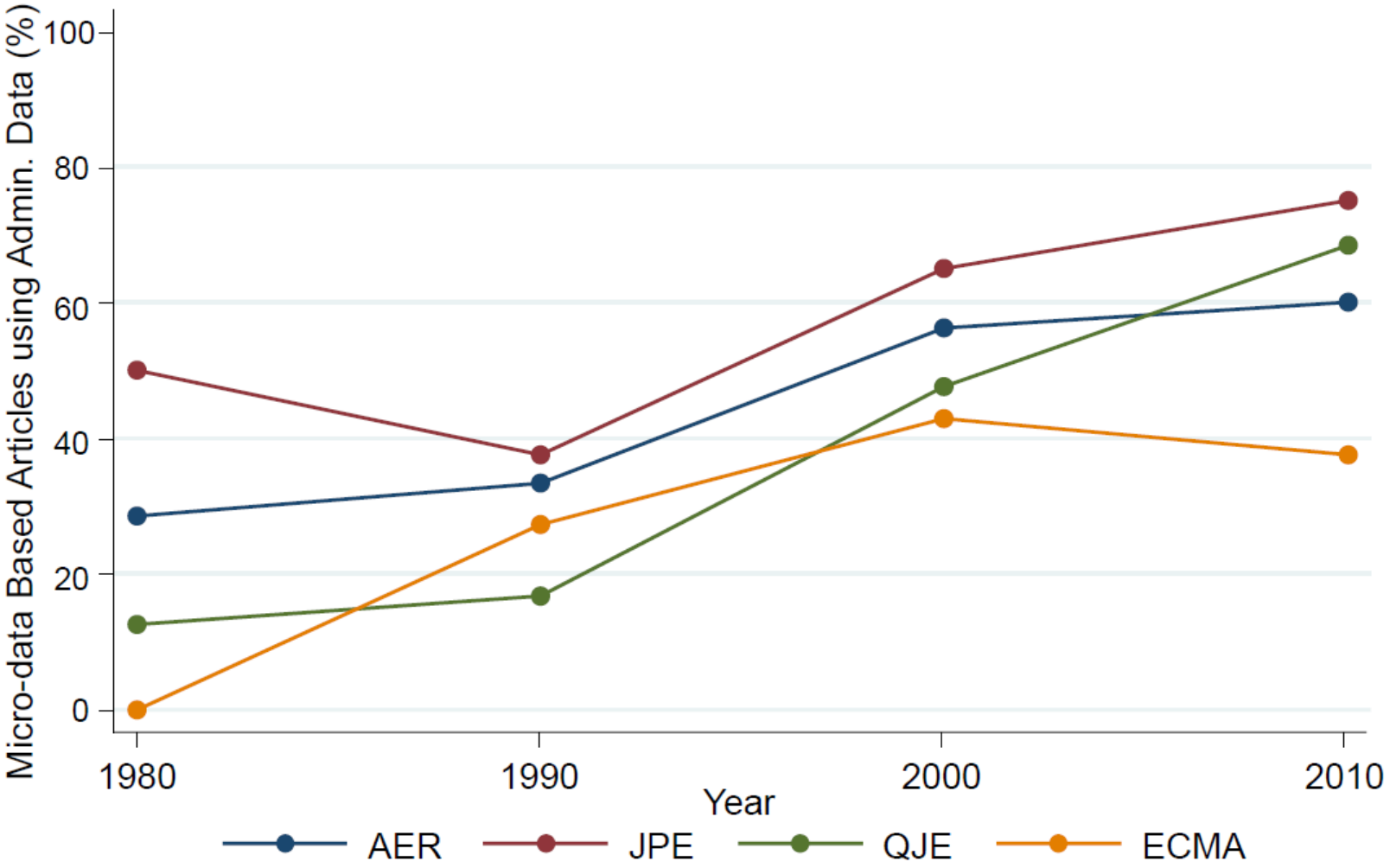
Use of Pre-Existing Survey Data in Publications in Leading Journals, 1980-2010



Note: "Pre-existing survey" datasets refer to micro surveys such as the CPS or SIPP and do not include surveys designed by researchers for their study. Sample excludes studies whose primary data source is from developing countries.

Source: Raj Chetty, Harvard University

Use of Administrative Data in Publications in Leading Journals, 1980-2010



Note: "Administrative" datasets refer to any dataset that was collected without directly surveying individuals (e.g., scanner data, stock prices, school district records, social security records). Sample excludes studies whose primary data source is from developing countries.

Source: Raj Chetty, Harvard University

Record Linkage*

HALBERT L. DUNN, M.D., F.A.P.H.A.

Chief, National Office of Vital Statistics, U. S. Public Health Service, Federal Security Agency, Washington, D. C.

EACH person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this Book into a volume.

The Book has many pages for some and is but a few pages in length for others. In the case of a stillbirth, the entire volume is but a single page.

The person retains the same identity throughout the Book. Except for ad- the various important records of a person's life.

The two most important pages in the Book of Life are the first one and the last one. Consequently, in the process of record linkage the uniting of the fact-of-death with the fact-of-birth has been given a special name, "death clearance."

IMPORTANCE OF ASSEMBLING THE BOOK OF LIFE

There are many uses for the important records of each person, brought to-

Traditional linkage

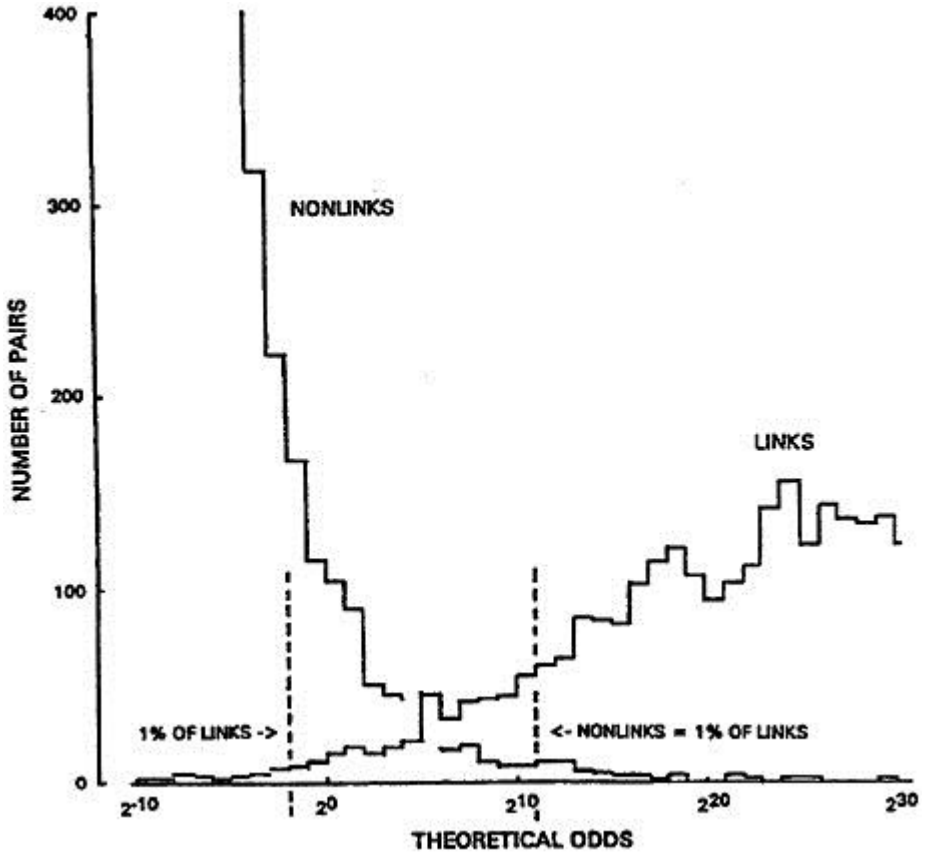
- Determinist
- Probabilistic

Traditional linkage

- Determinist

Traditional linkage

- Determinist
- Probabilistic




THE HISTORY KEEPERS
THE STORM BEGINS
DAMIAN DIBBEN
THE HISTORY KEEPERS
CIRCUS MAXIMUS
DAMIAN DIBBEN
THE HISTORY KEEPERS
NIGHTSHIP TO CHINA
DAMIAN DIBBEN

Multiple-entity resolution

Socioeconomic disadvantage, fetal environment and child development: linked Scottish administrative records based study
Playford, C. J., Dibben, C. & Williamson, L. 22 Nov 2017 In : International Journal for Equity in Health. 16, 1, p. 203 1 p.
Research output: Contribution to journal › Article

Maternal exposure to ambient air pollution and fetal growth in North-East Scotland: A population-based study using routine ultrasound scans
Clemens, T., Turner, S. & Dibben, C. Oct 2017 In : Environment International. 107, p. 216-226 10 p.
Research output: Contribution to journal › Article

A cancer geography paradox?: Poorer cancer outcomes with longer travelling times to healthcare facilities despite prompter diagnosis and treatment: a data-linkage study
Turner, M., Fielding, S., Ong, Y., Dibben, C., Feng, Z., Brewster, D. H., Black, C., Lee, A. & Murchie, P. 22 Jun 2017 In : British Journal of Cancer.
Research output: Contribution to journal › Article

Practical Data Synthesis for Large Samples
Raab, G., Nowok, B. & Dibben, C. 31 May 2017 In : Journal of Privacy and Confidentiality. 7, 3, 4
Research output: Contribution to journal › Article

Does equality legislation reduce intergroup differences? Religious affiliation, socio-economic status and mortality in Scotland and Northern Ireland: A cohort study of 400,000 people
Wright, D. M., Rosato, M., Raab, G., Dibben, C., Boyle, P. & O'reilly, D. 1 May 2017 In : Health and Place. 45, p. 32-38
Research output: Contribution to journal › Article

Pregnancy outcome and ultraviolet radiation; A systematic review
Megaw, L., Clemens, T., Dibben, C., Weller, R. & Stock, S. May 2017 In : Environmental Research. 155, p. 335-343 9 p.
Research output: Contribution to journal › Article

The effect of ultraviolet radiation on birth weights and gestational length in a scottish birth cohort
Clemens, T., Lauren, M., Dibben, C., Stock, S. & Weller, R. 18 Apr 2017 In : International Journal for Population Data Science. 1, 1
Research output: Contribution to journal › Article

Scottish vital events

- Civil registration of births, deaths and marriages in Scotland began on 1 January 1855

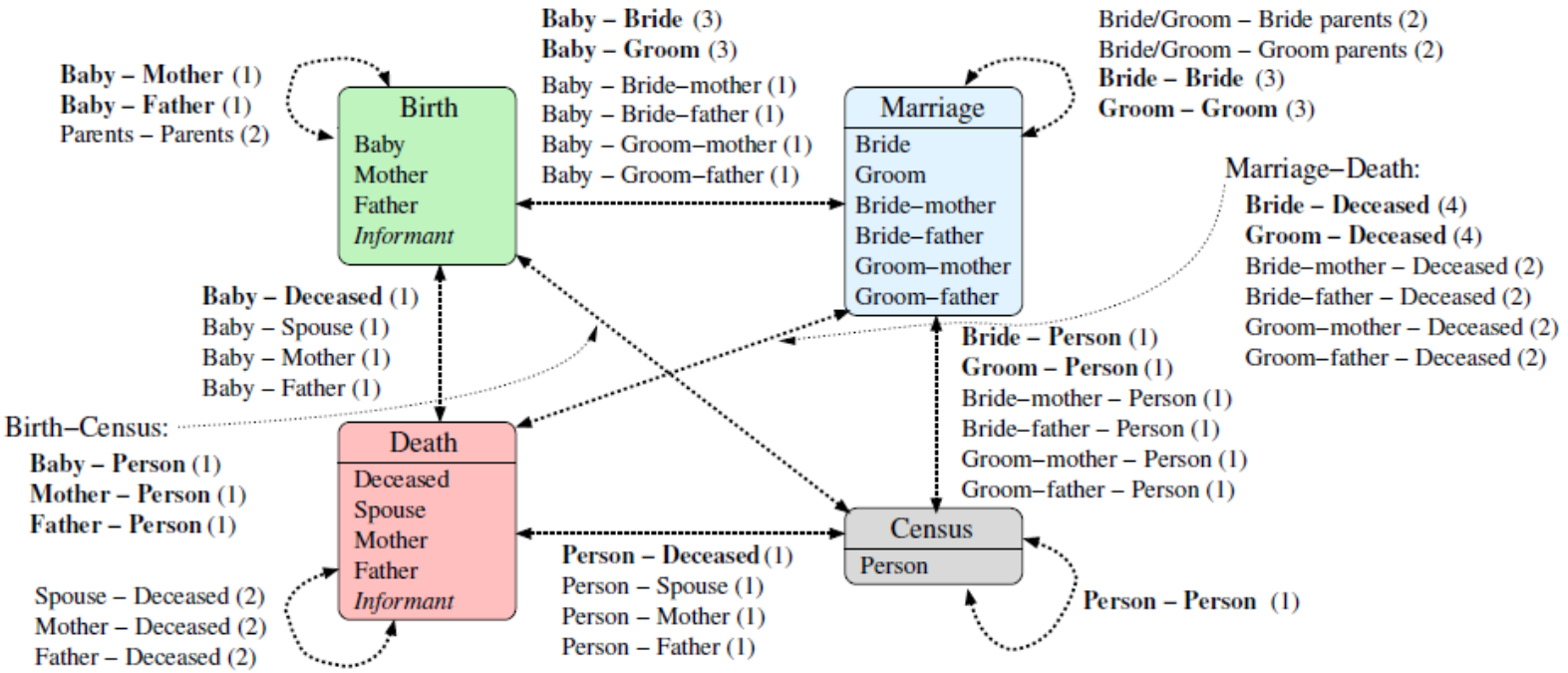


Fig. 1. Structure of the domain under consideration. The given numbers show how many individuals per certificate can be linked for a certain pair of roles. Pairs in bold font are those we aim to link for life segments such as the ones shown in Fig. 2.

Research Article

Statistics in Medicine

Received 21 June 2011 Accepted 10 June 2012 Published online 17 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5508

The analysis of record-linked data using multiple imputation with data value priors

Harvey Goldstein,[a,b*†] Katie Harron[a] and Angie Wade[a]

Probabilistic record linkage techniques assign match weights to one or more potential matches for those individual records that cannot be assigned 'unequivocal matches' across data files. Existing methods select the single record having the maximum weight provided that this weight is higher than an assigned threshold. We argue that this procedure, which ignores all information from matches with lower weights and for some individuals assigns no match, is inefficient and may also lead to biases in subsequent analysis of the linked data. We propose that a multiple imputation framework be utilised for data that belong to records that cannot be matched unequivocally. In this way, the information from all potential matches is transferred through to the analysis stage. This procedure allows for the propagation of matching uncertainty through a full modelling process that preserves the data structure. For purposes of statistical modelling, results from a simulation example suggest that a full probabilistic record linkage is unnecessary and that standard multiple imputation will provide unbiased and efficient parameter estimates.

Keywords: linking errors; missing data; multiple imputation; prior informed imputation; record linkage

Record	Set *A* variables		Set *B* variables		
1	0	0	*X*	*X*	*X*
2	0	0	*X*	0	*X*
3	0	0	*X*	*X*	0
4	0	0	0	0	*X*

Figure 6.1 Primary data file with four records where the set *B* variables are recorded and the set *A* variables are located in a linking file. *X* represents a recorded variable value and 0 a missing value. Initially, all the set *A* variables are missing and also some set *B* variables are missing as shown.

Record	Set *A* variables		Set *B* variables		
1	0	0	*X*	*X*	*X*
2	*X*	*X*	*X*	0	*X*
3	0	*X*	*X*	*X*	0
4	0	0	0	0	*X*

Figure 6.2 Primary data file with four records where the primary record file set *B* variables are recorded and the set *A* variable values for records 2 and 3 have been correctly transferred, unequivocally, via deterministic linkage with a linking file. *X* represents a variable with known value and 0 a missing value.

7

Using graph databases to manage linked data

James M. Farrow
SANT Datalink, Adelaide, South Australia, Australia
Farrow Norris, Sydney, New South Wales, Australia

7.1 Summary

Linked data has traditionally been managed using relational databases and methodologies that for historical reasons have been optimised to minimise the use of memory and persistent storage. This approach discourages exploration of the relationship between linked records because such information is either not retained or, depending on how it is stored, is difficult to exploit.

Linked data naturally form a *graph* or *network*: a collection of *nodes* (the records)

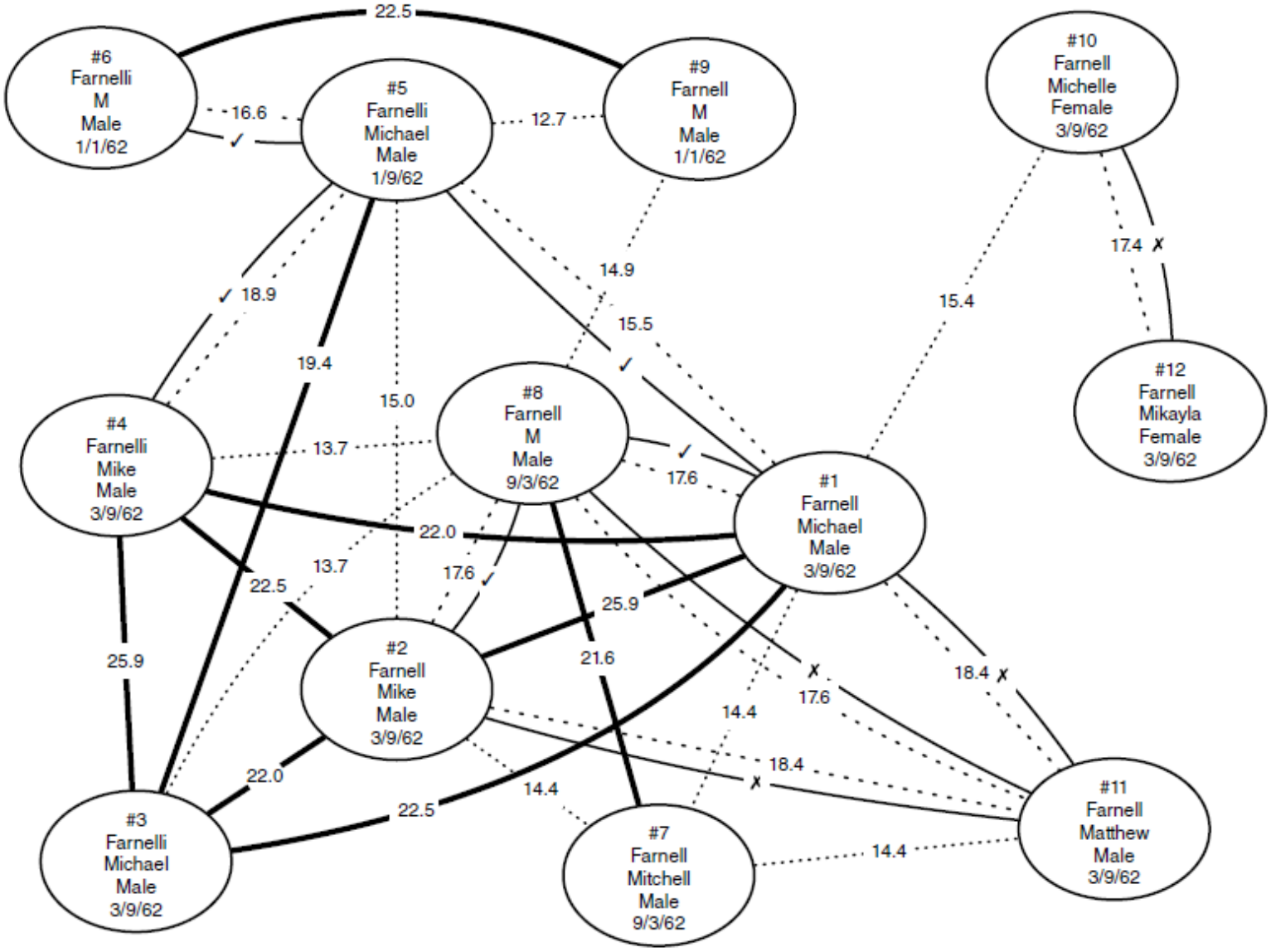


Figure 7.12 Comparison graph with review links.

‘End to end’ approach

- Rather than a final single ‘correct’ dataset
- Need to retain information from all aspects of the process

- Approving projects
- De-identified data
- Trusted researchers
- Secure environment
- Safe results
- Legal framework

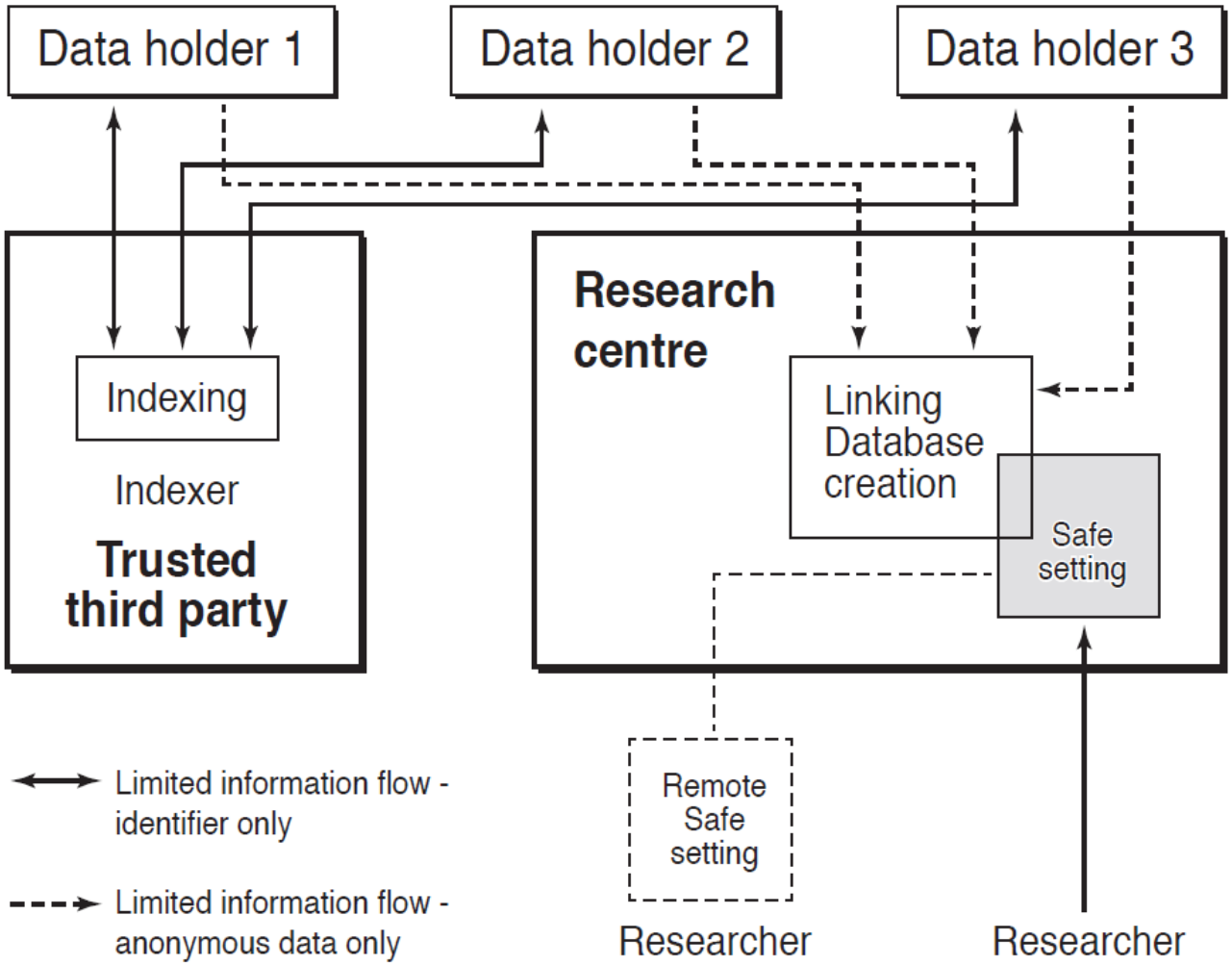
Protecting privacy

Share

We aim to have the highest standards of secure data sharing, which will be consistent across the Network. This means:

- safe projects - only projects approved by our Approvals Panel will have access to our services.
- safe people - only accredited researchers will have access to the Network's services.
- safe, de-identified data: researchers will not be able to see information which directly identify any individual.
- secure environments: state-of-the-art secure information technology and procedures will provide physical, hardware and software security across the whole Network.

Privacy-ethical advisory bodies



Conclusion

- Clear benefits from taking a more holistic approach to the process of data linkage
- However there are problems to overcome

Journal of Public Health | pp. 1–8 | doi:10.1093/pubmed/fdx037

Perspectives

GUILD: GUidance for Information about Linking Data sets†

Ruth Gilbert[1], Rosemary Lafferty[1], Gareth Hagger-Johnson[1], Katie Harron[2], Li-Chun Zhang[3], Peter Smith[3], Chris Dibben[4], Harvey Goldstein[1]

[1]Administrative Data Research Centre for England, University College London Great Ormond Street Institute of Child Health, London, UK
[2]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK
[3]Department of Social Statistics and Demography, University of Southampton, Southampton, UK
[4]Administrative Data Research Centre for Scotland, University of Edinburgh, Edinburgh, UK
Address correspondence to Ruth Gilbert, E-mail: r.gilbert@ucl.ac.uk.

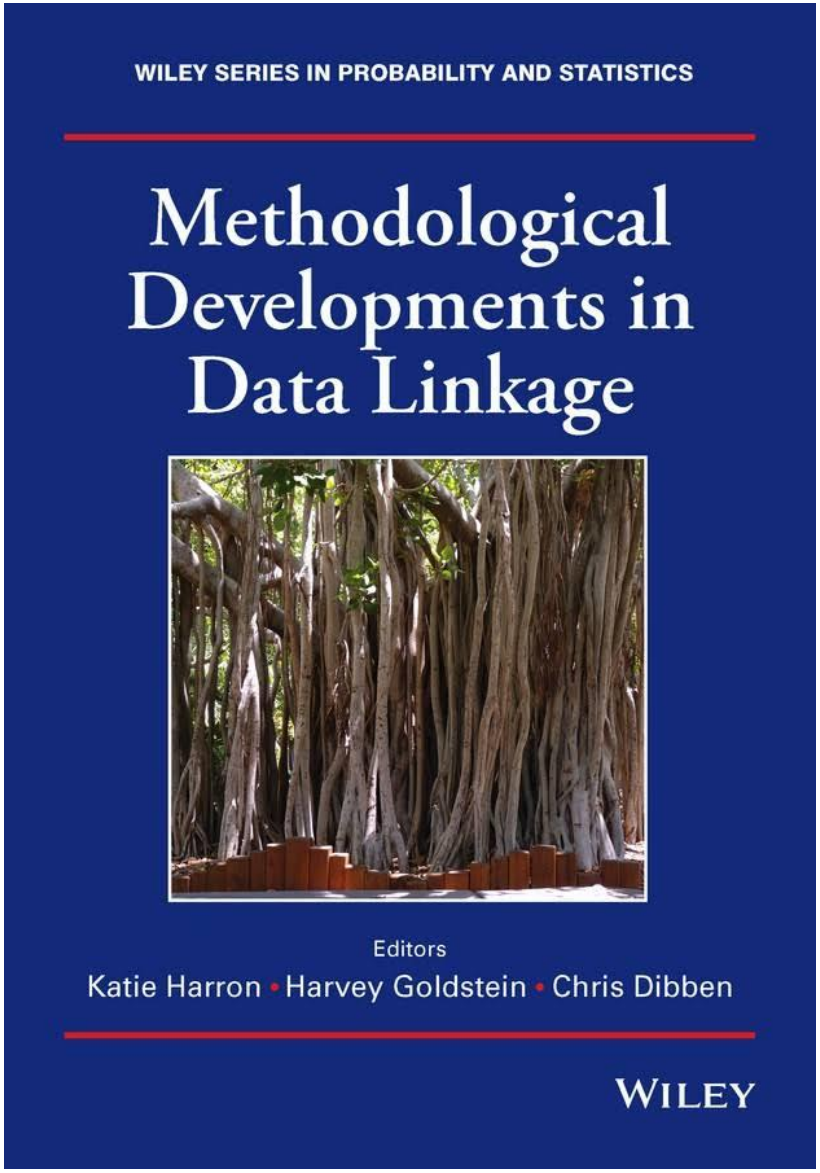
ABSTRACT

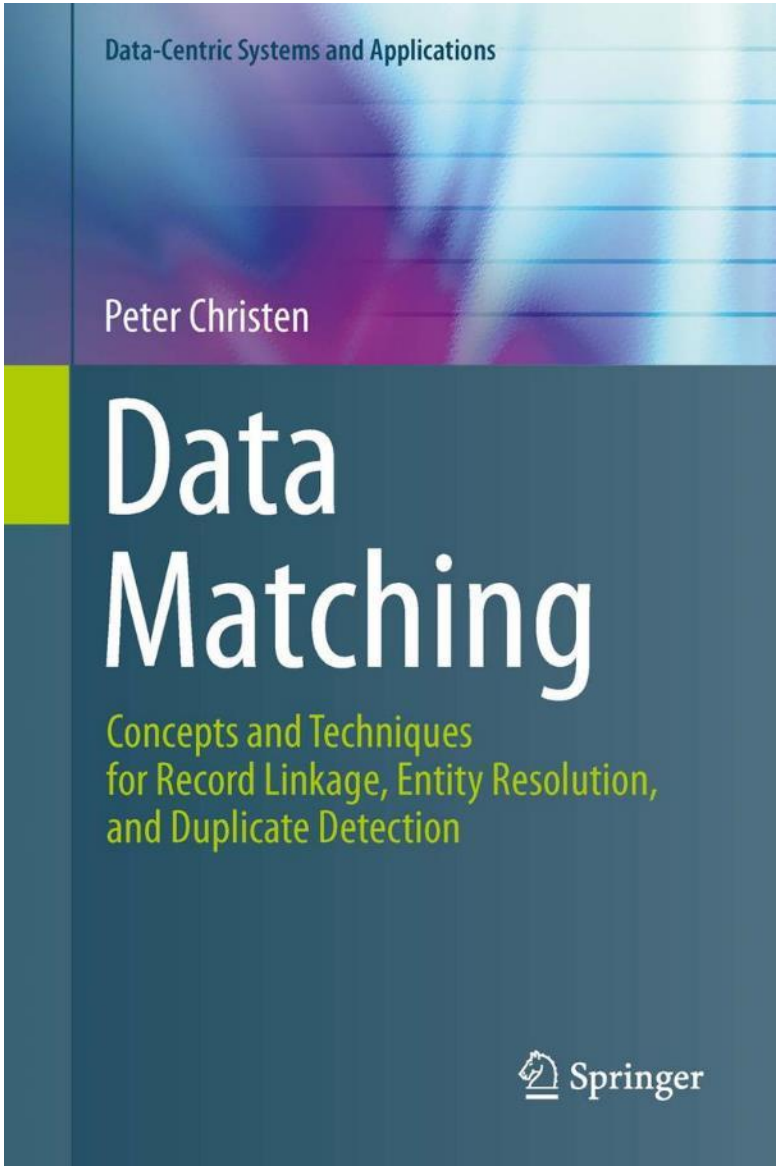
Record linkage of administrative and survey data is increasingly used to generate evidence to inform policy and services. Although a powerful and efficient way of generating new information from existing data sets, errors related to data processing before, during and after linkage can bias results. However, researchers and users of linked data rarely have access to information that can be used to assess these biases or take them into account in analyses. As linked administrative data are increasingly used to provide evidence to guide policy and services, linkage error, which disproportionately affects disadvantaged groups, can undermine evidence for public health. We convened a group of researchers and experts from government data providers to develop guidance about the information that needs to be made available about the data linkage process, by data providers, data linkers, analysts and the researchers who write reports. The guidance goes beyond recommendations for information to be included in research reports. Our aim is to raise awareness of information that may be required at each step of the linkage pathway to improve the transparency, reproducibility, and accuracy of linkage processes, and the validity of analyses and interpretation of results.

Keywords epidemiology, health services, management and policy

Table 1 GUILD guidance information to be shared before, during and after data linkage

Item	*Concept*	*Guidance*
Step 1	Data provision	
1a	Population included in the data set	Data providers should give details of the population included in the data set (e.g. everyone registered with a GP), the geographic coverage of the data (e.g. England and Wales), the number of records in each source data set and how any 'opt-outs' were dealt with
1b	Linkability of the data set	Details should be shared about how the data were generated (e.g. face-to-face), processed (e.g. a self-entered form or entered by an administrator) and quality controlled (e.g. manually checked), including how identifying characteristics were
1b(i)		– Collected and allocated
1b(ii)		– Updated as further personal data were collected, and dates of most recent updates
1b(iii)		– Checked and cleaned, including any validation rules
1b(iv)		– Replaced with artificial identifiers to reduce disclosure before being released for linkage
Step 2	Data linkage	
2a	Descriptions of linkage processes	Data linkers should provide descriptions of how the linkage was done including:
2a(i)		– A clear description of the data sources and identifying characteristics used for linkage, details of how identifiers were cleaned and validated before linkage, patterns of missingness, the expected range of values after cleaning, and how any de-duplication was performed.
2a(ii)		– Details of any transformation or replacement with artificial identifiers before linkage
2a(iii)		– A detailed description of the method (or algorithm) used for linkage, whether it was rule-based (e.g.


WILEY SERIES IN PROBABILITY AND STATISTICS
Methodological Developments in Data Linkage
Editors
Katie Harron • Harvey Goldstein • Chris Dibben
WILEY


Data-Centric Systems and Applications
Peter Christen
Data Matching
Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection
Springer