Linking administrative data for research

Katie Harron, Sir Henry Wellcome Postdoctoral Fellow
Department of Health Services Research and Policy, LSHTM
February 2017
A statistical definition

“a merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record”

Organisation for Economic Co-operation and Development (OECD)
Glossary of Statistical Terms
Record linkage for health data

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.

Dunn, 1946
Dunn applies "Record Linkage" to health research.

Fellegi and Sunter formalise probabilistic linkage.

The Oxford Record Linkage System,

The Western Australia Data Linkage System,

The Scottish Record Linkage System,

The Manitoba Population-Based Health Information System,

MRC and ESRC fund E-health and Admin data centres in UK,

SAIL databank established in Wales,

Newcombe proposes using probabilities in linkage.

Howard Newcombe
Opportunities and challenges linking administrative data

**Opportunities**

+ population-level resource
+ (potentially) lower risk of selection bias - generalisability
+ allows evaluation of rare events / hard to reach subgroups
+ detailed longitudinal trajectories
+ cost effective – exploits existing data
+ answer novel research questions
Deep vein thrombosis and air travel: record linkage study

C W Kelman, M A Kortu, N G Becker, Z Li, J D Mathews, C S Guest, C D J Holman

Abstract

Objective To investigate the time relations between pulmonary embolism after long flights has brought the issue to public attention. The incidence of venous thromboembolism varies

Conclusions The annual risk of venous thromboembolism is increased by 12% if one long haul flight is taken yearly.

Electronic data on flight arrivals and departures

Hospitalisations data

Conclusions The annual risk of venous thromboembolism is increased by 12% if one long haul flight is taken yearly. The average risk of death from flight related venous thromboembolism is small compared with that from motor vehicle crashes and injuries at work. The individual risk of death from flight related venous thromboembolism for people with certain pre-existing medical conditions is, however, likely to be greater than the average risk of 1 per 2 million for passengers arriving from a flight. Airlines and health authorities should continue to advise passengers on how to minimise risk.

Participants and methods

Data included coded personal identifiers, age, sex, arrival and departure dates, and nationality of the trav-
Answering novel research questions

Figure 3.3 Raw differences in conception behaviour by the end of Year 11, by Key Stage 2 English scores: estimated odds ratios relative to attaining Level 4

CAYT Impact Study: Report No. 6

Claire Crawford
Jonathan Cribb
Elaine Kelly
Opportunities and challenges linking administrative data

**Opportunities**

+ population-level resource
+ (potentially) lower risk of selection bias - generalisability
+ detailed longitudinal healthcare trajectories
+ allows evaluation of rare events / hard to reach subgroups
+ cost effective – exploits existing data
+ answer novel research questions

**Challenges**

- uncertainty about data quality
- lack of unique identifiers for linkage
- data security considerations
Identifiers
Trial ID
Clinical data

Linkage unit
(e.g. NHS Digital)
Assigns linkage key

Research group
Linkage Key
(Trial ID + Pupil ID)

Linked anonymised data

Data custodian 1
e.g. CTU

Identifiers
Trial ID
Clinical data

Data custodian 2
e.g. DfE

Identifiers
Pupil ID
Education data

Identifiers
Pupil ID

Clinical data
Trial ID

Identifiers
Trial ID
Linking hospital records for mothers and babies

Opportunities

- novel research questions, involving rare outcomes
  → induction of labour and perinatal mortality / neonatal morbidity
  → maternal mortality following neonatal abstinence syndrome

Challenges

- uncertainty about data quality
  → can we use linkage to better understand / improve data quality?
- lack of unique identifiers for linkage
  → how can we handle bias due to linkage error?
Hospital Episode Statistics

Postcode district, Ethnicity, GP practice, Provider

Episode dates
Diagnoses (ICD10)
Operations (OPCS)

Mother | Main record

Baby (card record) | Main record

Postcode district, Ethnicity, GP practice, Provider

Episode dates
Diagnoses (ICD10)
Operations (OPCS)

Hospital Episode Statistics

Linkage methods

• **Deterministic**
  - Rule-based approach, often looking for exact agreement on a number of identifiers

• **Probabilistic**
  - Uses the conditional probability that identifiers on different records will agree
    • Given records belong to the same person
    • Given records belong to different people (~ agreement by chance)
Probabilistic linkage

pair 1
Low match weight

pair 2

pair 3
High match weight

Baby File
Liza Minnelli

Mother File
Judy Garland
Low match weight

disagreement on date of birth

High match weight

agreement on NHS number

agreement on sex

\[ P(\gamma=1 \mid M) = m\text{-probability} \]

the probability of agreement given the records from same subject

\[ P(\gamma=1 \mid U) = u\text{-probability} = \]

the probability of agreement given the records from different subjects

Log ratio = \( w \) =

\[
\log_2 \left( \frac{m}{u} \right) \\
\log_2 \left[ \frac{(1-m)}{(1-u)} \right]
\]

if identifiers agree

if identifiers disagree

Match weight = \( W = \sum w_i \)
True matches

Non-matches

Low match weight

- Disagreement on identifiers
  - Chance (same date of birth)
  - Missing data

- Recording errors

High match weight

Agreement on identifiers
High specificity

Probabilistic linkage

Low match weight

Non-matches

High sensitivity

Missed matches

False matches

High match weight

True matches

Manual review

Manual review
Linkage

Baby records 2012
N = 673,055

Maternal records 2012
N=671,436

Deterministic linkage:
- GP practice
- Maternal age
- Birth weight
- Gestation
- Birth order
- Sex of baby

280,939 linked baby records (42%)

391,705 remaining unlinked baby records

Probabilistic linkage

Clinical variables
- First antenatal assessment date
- Estimated delivery date
- Gestation at first antenatal assessment
- Delivery place (actual)
- Delivery place (intended)
- Delivery method
- Method to induce labour
- Anaesthetic given during labour or delivery
- Anaesthetic given post labour or delivery
- Status of person conducting delivery
- Resuscitation method
- Birth status
- Number of babies
- Episode start date
- Episode end date

Combining deterministic and probabilistic:
- 660,401 linked baby-mother records (98% of babies)

Partial Identifiers
- Postcode
- District
- Ethnicity

Probabilistic match weights

<table>
<thead>
<tr>
<th>A</th>
<th>GP practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Postcode district</td>
</tr>
<tr>
<td>C</td>
<td>Estimated delivery date</td>
</tr>
<tr>
<td>D</td>
<td>First antenatal assessment</td>
</tr>
<tr>
<td>E</td>
<td>Episode end</td>
</tr>
<tr>
<td>F</td>
<td>Birth weight</td>
</tr>
<tr>
<td>G</td>
<td>Episode start</td>
</tr>
<tr>
<td>H</td>
<td>Delivery place (intention)</td>
</tr>
<tr>
<td>I</td>
<td>Status of person conducting delivery</td>
</tr>
<tr>
<td>J</td>
<td>Maternal age</td>
</tr>
<tr>
<td>K</td>
<td>Ethnic group</td>
</tr>
<tr>
<td>L</td>
<td>Gestation at first antenatal visit</td>
</tr>
<tr>
<td>M</td>
<td>Gestational age</td>
</tr>
<tr>
<td>N</td>
<td>Anaesthetic during delivery</td>
</tr>
<tr>
<td>O</td>
<td>Method of delivery</td>
</tr>
<tr>
<td>P</td>
<td>Method to induce labour</td>
</tr>
<tr>
<td>Q</td>
<td>Anaesthetic post-delivery</td>
</tr>
<tr>
<td>R</td>
<td>Sex</td>
</tr>
<tr>
<td>S</td>
<td>Delivery place</td>
</tr>
<tr>
<td>T</td>
<td>Resuscitation method</td>
</tr>
<tr>
<td>U</td>
<td>Birth status</td>
</tr>
<tr>
<td>V</td>
<td>Number of babies</td>
</tr>
<tr>
<td>W</td>
<td>Birth order</td>
</tr>
</tbody>
</table>
Understanding data quality

Mother (delivery record) | Main record | Baby tail

Gestation complete in 84% → Preterm birth rate = 6.3%

With linkage of information from baby record:

Completeness of gestation increases: from 84% → 92%

Preterm birth rate increases: from 6.3% → 6.7%
Understanding data quality

ICD10: Z371 single still birth
Z373 twins, one live on still
Z374 twins, both stillborn
Z377 other multiple, stillborn
O364 maternal care for intrauterine death

Birth status: (live or still)

0.55%

0.49%

Mother (delivery record) | Main record | Baby tail
Understanding data quality

ICD10: Z371 single still birth
Z373 twins, one live on still
Z374 twins, both stillborn
Z377 other multiple, stillborn
O364 maternal care for intrauterine death

Birth status: (live or still)

<table>
<thead>
<tr>
<th>Mother (delivery record)</th>
<th>Main record</th>
<th>Baby tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live</td>
<td>Live</td>
<td>Still</td>
</tr>
<tr>
<td>99.34%</td>
<td>0.17%</td>
<td>668,141</td>
</tr>
<tr>
<td>Still</td>
<td>0.12%</td>
<td>0.38%</td>
</tr>
<tr>
<td>667,797</td>
<td>3639</td>
<td>675,734</td>
</tr>
</tbody>
</table>

0.55%
0.49%
Understanding data quality

With linkage of information from baby record:

800/1558 stillbirth conflicts resolved by triangulating information held on mother/baby records

- Checking ICD10 codes, birth status, length of stay
- 0.1% of records unresolved

<table>
<thead>
<tr>
<th>Birth status</th>
<th>ICD</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Live</td>
<td>Still</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td>99.34%</td>
<td>0.17%</td>
<td>668,141</td>
<td></td>
</tr>
<tr>
<td>Still</td>
<td>0.12%</td>
<td>0.38%</td>
<td>3295</td>
<td></td>
</tr>
<tr>
<td></td>
<td>667,797</td>
<td>3639</td>
<td>675,734</td>
<td></td>
</tr>
</tbody>
</table>
Lack of reliable unique identifiers → Linkage error

<table>
<thead>
<tr>
<th>Link status</th>
<th>Match status</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td>Identified match</td>
<td>False match</td>
</tr>
<tr>
<td>Non-link</td>
<td>Missed match</td>
<td>Identified non-match</td>
</tr>
</tbody>
</table>
The linkage problem

- Small amounts of linkage error can result in substantially biased results

- **False matches**
  - introduce variability and weaken the association between variables – bias to the null

- **Missed matches**
  - reduce our sample size and result in a loss of power – potential selection bias
### Table 3. Hazard Ratios for the Association Between Ethnicity and Mortality Using Three Linkage Criteria, 1989-2002

<table>
<thead>
<tr>
<th>Ethnicity and nativity</th>
<th>Relaxed</th>
<th>NCHS cut-points</th>
<th>Tightened</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB Hispanic</td>
<td>1.24***</td>
<td><strong>0.97</strong></td>
<td><strong>0.78</strong>*</td>
</tr>
<tr>
<td>US NH White</td>
<td>ref</td>
<td>ref</td>
<td>ref</td>
</tr>
</tbody>
</table>

\*p < .10. **p < .05. ***p < .001
<table>
<thead>
<tr>
<th>Maternal factors</th>
<th>Matched pairs</th>
<th>ISC residuals</th>
<th>MDC residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (years)</td>
<td>n = 250,186</td>
<td>n = 2,596</td>
<td>n = 3,798</td>
</tr>
<tr>
<td>Married</td>
<td>29.6</td>
<td>28.9</td>
<td>30.0</td>
</tr>
<tr>
<td>Australian-born mother</td>
<td>72.6</td>
<td>77.9</td>
<td>75.7</td>
</tr>
<tr>
<td>Birth in private hospital</td>
<td>22.0</td>
<td>27.1</td>
<td>28.9</td>
</tr>
<tr>
<td>Caesarean delivery</td>
<td>23.1</td>
<td>20.7</td>
<td>28.9</td>
</tr>
<tr>
<td>Diabetes</td>
<td>4.4</td>
<td>3.2</td>
<td>4.8</td>
</tr>
<tr>
<td>Hypertension</td>
<td>7.1</td>
<td>7.9</td>
<td>8.3</td>
</tr>
<tr>
<td>Stillbirth&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.5</td>
<td>4.6</td>
<td>3.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baby factors</th>
<th>Matched pairs</th>
<th>ISC residuals</th>
<th>MDC residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birthweight (g)</td>
<td>n = 253,538</td>
<td>n = 1,570</td>
<td>n = 3,157</td>
</tr>
<tr>
<td>&lt;1000</td>
<td>0.4</td>
<td>0.8</td>
<td>4.4</td>
</tr>
<tr>
<td>1000–1999</td>
<td>1.7</td>
<td>3.9</td>
<td>7.9</td>
</tr>
<tr>
<td>2000–2999</td>
<td>18.5</td>
<td>22.5</td>
<td>27.8</td>
</tr>
<tr>
<td>3000–3999</td>
<td>66.9</td>
<td>59.9</td>
<td>48.8</td>
</tr>
<tr>
<td>4000–4999</td>
<td>12.4</td>
<td>12.1</td>
<td>10.5</td>
</tr>
<tr>
<td>≥5000</td>
<td>0.2</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Plurality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singletons</td>
<td>96.7</td>
<td>95.4</td>
<td>95.5</td>
</tr>
<tr>
<td>Twins</td>
<td>3.2</td>
<td>4.6</td>
<td>4.2</td>
</tr>
<tr>
<td>Death in hospital</td>
<td>0.2</td>
<td>0.9</td>
<td>2.8</td>
</tr>
<tr>
<td>Preterm birth&lt;sup&gt;b&lt;/sup&gt;</td>
<td>6.5</td>
<td>9.7</td>
<td>26.3</td>
</tr>
<tr>
<td>Transfer to another hospital</td>
<td>5.3</td>
<td>11.9</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Variation - errors in identifiers in HES

Variation - errors in identifiers in HES

Differential (non-random) linkage – why?

- Data quality differs by patient group / SES etc.
  - Bohensky et al 2010. Data Linkage: A powerful research tool with potential problems. *BMC Health Services Research*

- Unknown/estimated dates of birth
  - Unconscious, frail, dementia,

- Unconventional surnames

- Misleading information
  - Drug user, parent withholding details

- Address issues
  - Communal establishments
  - Visitor / tourist / traveller

- Multiple births
  - Same sex, postcode, date of birth
Evaluating linkage

i) Comparisons of linked and unlinked data

<table>
<thead>
<tr>
<th>Matched pairs</th>
<th>BSC residuals</th>
<th>MDC residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed factors</td>
<td>a = 250-186</td>
<td>a = 2506</td>
</tr>
<tr>
<td>Mean age (years)</td>
<td>29.6</td>
<td>29.9</td>
</tr>
<tr>
<td>Married</td>
<td>76.7</td>
<td>73.4</td>
</tr>
<tr>
<td>Age at death</td>
<td>76.8</td>
<td>73.4</td>
</tr>
<tr>
<td>Cause of death</td>
<td>25.1</td>
<td>21.9</td>
</tr>
<tr>
<td>Diabetes</td>
<td>4.4</td>
<td>5.2</td>
</tr>
<tr>
<td>Hypertension</td>
<td>7.1</td>
<td>7.9</td>
</tr>
<tr>
<td>Respiratory</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td>Birthweight (g)</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>&lt;1500</td>
<td>4.0</td>
<td>0.8</td>
</tr>
<tr>
<td>1500-1999</td>
<td>1.7</td>
<td>3.9</td>
</tr>
<tr>
<td>2000-2499</td>
<td>16.5</td>
<td>22.9</td>
</tr>
<tr>
<td>2500-2999</td>
<td>68.9</td>
<td>56.9</td>
</tr>
<tr>
<td>3000-3999</td>
<td>12.4</td>
<td>12.1</td>
</tr>
<tr>
<td>4000</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Plurality</td>
<td>96.7</td>
<td>95.4</td>
</tr>
<tr>
<td>Singletons</td>
<td>3.2</td>
<td>4.6</td>
</tr>
<tr>
<td>Twins</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Death in hospital</td>
<td>5.3</td>
<td>5.3</td>
</tr>
<tr>
<td>Hospital stay</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Transfer to another hospital</td>
<td>3.3</td>
<td>13.9</td>
</tr>
</tbody>
</table>

Highly sensitive

Highly specific

ii) Gold-standard / reference data to quantify linkage errors

iii) Sensitivity analysis using different probabilistic thresholds
Evaluating mother-baby linkage
Evaluating mother-baby linkage

Maternal Age

Deprivation quintile (IMD)

Birth weight

Gestation

Linked

Unlinked

98%

2%
Evaluating mother-baby linkage

Gold-standard: 15 maternity units, 2012/13 (N=72,824)

Original algorithm
- 632 (0.9%) false matches
- 297 (0.4%) missed matches

Sensitivity analysis: Different linkage criteria

Conservative prob algorithm
- 212 (0.3%) false matches
- 7,797 (10.7%) missed matches

Deterministic only algorithm
- 22 (0.1%) false matches
- 37,515 (51.6%) missed matches
Evaluating mother-baby linkage

Identifying impact on results: rate of preterm birth

% of births <37 weeks of gestation

- Gold-standard
- Probabilistic
- Conservative
- Deterministic

Values:
- Gold-standard: 6.8
- Probabilistic: 7.2
- Conservative: 6.4
- Deterministic: 6.6
Evaluating mother-baby linkage

Identifying impact on results: association between maternal risk-factors and infant survival to discharge
Alternative linkage methods

Attribute-specific match weights
- Exploiting what we know about variation in identifier errors
- E.g. when errors in identifiers differ between ethnic groups

![Bar chart showing % readmitted for different ethnic groups (Truth vs. Traditional)]
Alternative linkage methods

Attribute-specific match weights
- Exploiting what we know about variation in identifier errors
- E.g. when errors in identifiers differ between ethnic groups

Alternative linkage methods

Calculation of match weights/scores without the need for training data

– Class of correspondence analysis
– Set constraints, e.g.
  • full agreement = 100,
  • no agreement = 0
– Given a set number of identifiers and levels of agreement between those identifiers, aim is to derive scores that minimise the total discrepancy within each pair of records

<table>
<thead>
<tr>
<th></th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaling scores</td>
<td>53</td>
<td>22</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>Probabilistic weights</td>
<td>32</td>
<td>27</td>
<td>26</td>
<td>15</td>
</tr>
</tbody>
</table>

## Accounting for linkage error

<table>
<thead>
<tr>
<th>Baby record</th>
<th>Maternal record</th>
<th>Risk factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact match</td>
<td>Yes</td>
</tr>
<tr>
<td>Record 1</td>
<td>Exact match</td>
<td>Yes</td>
</tr>
<tr>
<td>Record 2</td>
<td>Exact match</td>
<td>No</td>
</tr>
<tr>
<td>Record 3</td>
<td>Match weight=10</td>
<td>Yes</td>
</tr>
<tr>
<td>Record 4</td>
<td>Match weight=8</td>
<td>No</td>
</tr>
<tr>
<td>Record n</td>
<td>Match weight=5</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Match weight=4</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Match weight=4</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Match weight=3</td>
<td>Yes</td>
</tr>
</tbody>
</table>

### Traditional probabilistic

| Yes | Yes | No |

---

Goldstein et al *Stat Med* 2012;31(28):3481-3493

Imputation for missed links

<table>
<thead>
<tr>
<th>Record 1</th>
<th>Baby record</th>
<th>Maternal record</th>
<th>Risk factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Exact match</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exact match</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exact match</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Match weight=10</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Match weight=8</td>
<td>No</td>
</tr>
</tbody>
</table>

Multiple imputation

- Yes
- Yes
- No
- Yes

Goldstein et al *Stat Med* 2012;31(28):3481-3493
## Imputation for linkage uncertainty

<table>
<thead>
<tr>
<th>Baby record</th>
<th>Maternal record</th>
<th>Risk factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Record 1</th>
<th>Exact match</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record 2</td>
<td>Exact match</td>
<td>Yes</td>
</tr>
<tr>
<td>Record 3</td>
<td>Exact match</td>
<td>No</td>
</tr>
<tr>
<td>Record 4</td>
<td>Match weight=10</td>
<td>Yes</td>
</tr>
<tr>
<td>Record n</td>
<td>Match weight=8</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Match weight=5</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Match weight=4</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Match weight=4</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Match weight=3</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Prior-informed imputation

Summary

• Linkage can help to address data quality issues
  – Improve ascertainment of key risk-factors and outcomes
  – Triangulate outcomes and resolve inconsistencies
  – Highlights limitations in the data

• Understanding bias due to linkage error is important
  – Several approaches available for evaluating potential impact on results
  – Requires information on linkage process and unlinked records (difficult with trusted third party model)

• Unfulfilled opportunities
  – Linkage between health and other sectors
  – Linkage of trial data for long-term follow up / safety analyses
Acknowledgements and funding

Fellowship steering committee:
Jan van der Meulen,
Ruth Gilbert
David Cromwell
Astrid Guttmann
Harvey Goldstein

Thanks also to Hannah Knight and Ipek Gurol (Royal College of Obstetricians and Gynecologists)

This work was supported by funding from the Wellcome Trust (103975/Z/14/Z)
Hospital Episode Statistics were made available by the NHS Health and Social Care Information Centre (Copyright © 2012, Re-used with the permission of The Health and Social Care Information Centre. All rights reserved.)