Enhancing observational data through linkage: opportunities and challenges

Dr Katie Harron, Sir Henry Wellcome Postdoctoral Fellow
London School of Hygiene and Tropical Medicine
Data linkage

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
Opportunities and challenges using administrative / clinical / routine data

+ population-level resource
+ detailed longitudinal healthcare trajectories
+ allows evaluation of rare events / specific subgroups
+ potentially lower risk of selection bias
+ 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 ranges.

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
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 using administrative / clinical / routine data

+ population-level resource
+ detailed longitudinal healthcare trajectories
+ allows evaluation of rare events / specific subgroups
+ potentially lower risk of selection bias
+ answer novel research questions

- uncertainty about data quality
- different ways to code the same outcome
- information found across a number of fields
- sometimes lack of consistency
Linkage to overcome data quality issues

+ Allows triangulation of outcomes
+ Improves ascertainment

Example 1:
Mother-baby linkage in English hospital data

Example 2:
Evaluating error in linkage of intensive care / laboratory data

- Complete, accurate identifiers not always available
- Potential for introducing bias due to linkage error
Mother-baby linkage in NHS

- Mother and baby records not routinely linked in data within the English National Health Service (NHS)

- Linked maternal-baby data is available in other countries
  - e.g. Scotland, Canada, Australia, Netherlands, US

- Linkage of prospective data planned for future in England

We evaluated whether linkage of retrospective data from maternal and baby records could be used to address data quality issues in English hospital data

- Deterministic and probabilistic linkage of “indirect” identifiers

- Non-disclosive variables contained in both maternal and baby records
Example: mother-baby linkage

- Episode dates
- Diagnoses (ICD10)
- Operations (OPCS)

- Delivery information
- Birth weight
- Gestation

- Postcode district,
  Ethnicity, GP practice, Provider

- Mother (delivery record) | Main record | Baby tail

- Baby (birth record) | Main record | Baby tail

- Episode dates
- Diagnoses (ICD10)
- Operations (OPCS)

- Delivery information
- Birth weight
- Gestation

- Postcode district,
  Ethnicity, GP practice, Provider

Example: mother-baby linkage
Example: mother-baby linkage

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

<table>
<thead>
<tr>
<th>Mother (delivery record)</th>
<th>Main record</th>
<th>Baby tail 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baby 1 (birth record)</td>
<td>Main record</td>
<td>Baby tail 1</td>
</tr>
<tr>
<td>Baby 2 (birth record)</td>
<td>Main record</td>
<td>Baby tail 2</td>
</tr>
</tbody>
</table>

Can linkage with baby records help improve ascertainment?
Example: mother-baby linkage

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%

Mother (delivery record) | Main record
-------------------------|------------------
Baby tail 1
Baby tail 2

Baby 1 (birth record) | Main record
----------------------|------------------
Baby tail 1

Baby 2 (birth record) | Main record
----------------------|------------------
Baby tail 2
Example: mother-baby linkage

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)

Can linkage with baby records help resolve inconsistencies?
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

380,164 linked baby records (96%)

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

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

Partial Identifiers:
- Postcode district
- Ethnicity
Combining information from baby and mother records

Ascertainment

• Completeness of gestation increases from 84% → 92%
• Preterm birth rate increases from 6.1% → 6.7%
  – Further increases to 6.9% using ICD10 code for preterm birth O60 in baby record

Inconsistencies

• 800/1558 stillbirth conflicts resolved through information held on baby record
  – Checking ICD10 codes, birth status, length of stay
  – 0.1% of records unresolved
Checking external validity

### Birth weight

#### Maternal age

<table>
<thead>
<tr>
<th>ONS</th>
<th>Linked HES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Still birth</td>
<td>0.53%</td>
</tr>
<tr>
<td>Multiple birth</td>
<td>3.2%</td>
</tr>
<tr>
<td>Preterm birth</td>
<td>7.1%</td>
</tr>
</tbody>
</table>
Comparing linked (98%) vs unlinked (2%)
<table>
<thead>
<tr>
<th>Link status</th>
<th>Match status</th>
<th>Match status description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td>Match (pair from same individual)</td>
<td>Identified match</td>
</tr>
<tr>
<td></td>
<td>Non-match (pair from different individuals)</td>
<td>False match</td>
</tr>
<tr>
<td>Non-link</td>
<td>Match (pair from same individual)</td>
<td>Missed match</td>
</tr>
<tr>
<td></td>
<td>Non-match (pair from different individuals)</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>0.97</td>
<td>0.78***</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

Lariscy. Differential Record Linkage by Hispanic Ethnicity and Age in Linked Mortality Studies: Implications for the Epidemiologic Paradox *(2011, J Aging Health 2011)*
<table>
<thead>
<tr>
<th></th>
<th>Matched pairs</th>
<th>ISC residuals</th>
<th>MDC residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maternal factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 250,186$</td>
<td>$n = 2596$</td>
<td>$n = 3798$</td>
</tr>
<tr>
<td>Mean age (years)</td>
<td>29.6</td>
<td>28.9</td>
<td>30.0</td>
</tr>
<tr>
<td>Married</td>
<td>78.7</td>
<td>73.4</td>
<td>NA</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>
<tr>
<td><strong>Baby factors</strong></td>
<td>$n = 253,538$</td>
<td>$n = 1570$</td>
<td>$n = 3157$</td>
</tr>
<tr>
<td>Birthweight (g)</td>
<td></td>
<td></td>
<td></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>

Differential linkage – why?

- Data quality differs by patient group:
  - 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

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

- Misleading information
  - Drug user, parent withholding details

- Multiple births
Evaluating bias due to linkage error

i) Subset of gold-standard data to quantify linkage bias

ii) Sensitivity analysis using different probabilistic thresholds

iii) Comparisons of linked and unlinked data

iv) Statistical / missing data methods - imputation for uncertain links
AIM: To evaluate trends in risk-adjusted infection rates in PICU
- Linkage using deterministic and probabilistic linkage
- Explored bias due to linkage error using
  - gold-standard data
  - sensitivity analyses
  - imputation

Example: Bloodstream infection in paediatric intensive care units

PICANet
(Paediatric Intensive Care Audit Network)
Admissions to Paediatric Intensive Care

LabBase2
(Public Health England)
National infection surveillance system
Example: Bloodstream infection in paediatric intensive care

Not all infections occur in PICU

Not all PICU admissions have an infection

PICANet
(Paediatric Intensive Care Audit Network)
Admissions to Paediatric Intensive Care

LabBase2
(Public Health England)
National infection surveillance system

Comparing characteristics of linked and unlinked data not helpful in this context
Linkage

<table>
<thead>
<tr>
<th>PICU admissions</th>
<th>PICANet</th>
<th>Common identifiers</th>
<th>Infection surveillance</th>
<th>LabBase2 (PHE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admission + discharge dates</td>
<td>Length of Stay</td>
<td>Risk-factors</td>
<td>Specimen date</td>
<td>Organism</td>
</tr>
</tbody>
</table>

- NHS number (51%)
- Hospital number (83%)
- First name (22%)
- Surname (26%)
- Date of birth (93%)
- Sex (96%)
- Postcode (64%)
- Location (100%)

Deterministic linkage

Probabilistic match weights

- Traditional probabilistic thresholds approach
- Imputation for uncertain links

Compare results with gold-standard subset of data
Choosing a threshold weight
Evaluation: sensitivity analysis
# Imputation for uncertain links

<table>
<thead>
<tr>
<th>Infection record</th>
<th>Admission record</th>
<th>Length of stay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record 1</td>
<td>Exact match</td>
<td>&gt;48 h</td>
</tr>
<tr>
<td>Record 2</td>
<td>Exact match</td>
<td>&gt;48 h</td>
</tr>
<tr>
<td>Record 3</td>
<td>Exact match</td>
<td>&lt;48 h</td>
</tr>
<tr>
<td>Record 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Record n</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Goldstein et al *Stat Med* 2012;31(28):3481-3493
## Evaluation: gold-standard data

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Number of links identified</th>
<th>Infection rate (%)</th>
<th>Bias (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold-standard</td>
<td>426</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>Relaxed (5)</td>
<td>492</td>
<td>4.5</td>
<td>15.5</td>
</tr>
<tr>
<td>Conservative (10)</td>
<td>418</td>
<td>3.8</td>
<td>-1.9</td>
</tr>
<tr>
<td>Imputation</td>
<td>424</td>
<td>3.9</td>
<td>-0.5</td>
</tr>
</tbody>
</table>
Results

Average annual decrease of 13.2% (95% CI 11.8-14.6%)

Rates of 2.87 (95% CI 2.40-3.35) per 1000 bed-days in 2012
Reporting studies using linked data

• Readers should also understand
  – Any limitations of the data
    • e.g. still birth
  – Processes by which errors occur
    • e.g. ethnic group
  – Implications for analyses
    • e.g. potential selection bias

• Transparency and clear reporting helps interpretation
  – Can be a challenge to obtain information on linkage process
• RECORD initiative: reporting guidelines for studies conducted using routinely-collected health data
  – record-statement.org
• Addresses unique challenges associated with using data collected primarily for reasons other than research
• Specific section on data linkage
RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.

RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.
Reporting

![Venn Diagram]

- Disease registry myocardial infarction: 1488 (8%)
- Primary care myocardial infarction: 3188 (18%)
- Hospital data myocardial infarction: 1806 (10%)
- Overlap areas:
  - Primary care and disease registry: 1099 (6%)
  - Primary care and hospital data: 5561 (31%)
  - Disease registry and hospital data: 3532 (20%)

Total myocardial infarctions: 5561 (31%)
Reporting

RACE-ER STEMI patients who arrive at a PCI center by EMS (n = 8680)

PreMIS

Applied linkage

Linked (n = 6010)

Not linked (n = 2670)

Table I: Patients' characteristics comparison using RACE variables

Directly arrived to a PCI center (n = 3077)

Excluded:
1) Patients treated by lytic (n = 11)
2) Patients with incomplete treatment data (n = 381)

Arrived at a non-PCI hospital first then transferred to a PCI center (n = 2933)

Excluded:
1) Patients treated by lytic (n = 912; including 684 patients who were treated with both lytic and PCI)
2) Patients with incomplete treatment data (n = 271)

Directly arrived and treated by PCI only (n = 2696)

Table III: Comparison of key metrics between RACE and PreMIS calculations

Transferred from non-PCI and treated by PCI only (n = 1750)

Table II: Patients' characteristics comparison using RACE variables
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
  – Transparent reporting can aid interpretation

• Unfulfilled opportunities – linkage between health and other sectors on new scale
  – e.g. Brazil’s 100 million cohort: socio-economic and health data
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.)