Challenges and opportunities in using administrative data linkage for research: the importance of quality assessment for understanding bias

Katie Harron and James Doidge
UCL Great Ormond Street Institute of Child Health
January 2020
k.harron@ucl.ac.uk
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
Electronic / administrative data cohorts

- Population cohorts created entirely from linkage of administrative data sources
  - e.g., linkage of mothers and babies within hospital data, in England and beyond...

Non-health data →

- Social environment exposures
- Educational outcomes
Linking information on perinatal outcomes to maternity records

- RCT evidence suggests induction of labour at 39 weeks has no short-term adverse effect on mother/infant among nulliparous women aged 35 years or older.

- The trial was underpowered to address the effect of routine induction of labour on the risk of perinatal death.

Perinatal outcomes after induction of labour compared with expectant management at 40 weeks gestation

562 inductions of labour at 40 weeks would be required to prevent 1 perinatal death.

66% lower risk of perinatal death (0.08% versus 0.26%)
Supplementing mortality data with more complete information on risk factors

Linkage to maternal records increased completeness of risk factors:
- 67% to 84% for birth weight
- 64% to 78% for gestational age
- 63% to 97% for maternal age
- 45% to 97% for IMD

Was excess child mortality in England compared with Sweden explained by the unfavourable distribution of birth characteristics in England?

The coverage of the complete case cohort increased from 18% to 75% of all births in HES-ONS birth cohort.

Data/figure courtesy of Ania Zylbersztejn

Challenges

- Administrative data not designed for linkage
- Unique identifiers may not be present in all sources
- Requires appropriate linkage methods

- (Identifier) data quality
- Linkage errors

- False matches and missed matches
- Can lead to biased results
- Requires appropriate analysis methods
How is linkage done?

- Deterministic (rule-based)

1. Sex
   - Date of Birth
   - NHS Number

2. Sex
   - Date of Birth
   - Postcode
   - Local Patient Identifier within Provider

3. Sex
   - Date of Birth
   - Postcode
How is linkage done?

- Probabilistic linkage

- Agreement on NHS number

- Agreement on sex

- Disagreement on date of birth

= Match weight / score
Mother (delivery record) | Main record | Baby tail

Baby (birth record) | Main record | Baby tail

---

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

Match status (true relationship)

<table>
<thead>
<tr>
<th>Link status</th>
<th>Match</th>
<th>Non-match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link</td>
<td>True link</td>
<td>False link</td>
</tr>
<tr>
<td>Non-link</td>
<td>Missed link</td>
<td>True non-link</td>
</tr>
</tbody>
</table>

PPV  
Proportion of links that are true

Sensitivity  
Proportion of matches that are linked
High PPV, low sensitivity

Definite matches

Definite non-matches

Possible matches

Missed links

Non-matches

Matches

High agreement

Low agreement
Low agreement

Non-matches

Definite non-matches

Possible matches

Definite matches

High agreement

False links

Low PPV, high sensitivity

Matches
Bias due to linkage error

• Large body of evidence showing that even small amounts of linkage errors can introduce substantial bias to results
  – Particularly important when errors are non-random, or more likely to occur for particular subgroups

• Methods for handling bias due to linkage error have been highlighted as a priority for research
Classifying linkage designs

Impact depends on the linkage classification and the question you are asking...

- ‘Master’
- ‘Intersection’
- ‘Union’
- ‘Disjunctive union’
- ‘Set difference’
- ‘Perfect overlap’
Classifying linkage designs

Impact depends on the linkage classification and the question you are asking...

**Question 1: Is the linkage meaningfully interpreted?**

Yes – to define outcomes (e.g. cancer diagnoses)
- Link / no link = cancer / cancer free

Yes – to define study population (e.g. children with Down’s syndrome)
- Link / no link = included / excluded from sample

No – to add other information (e.g. socio-demographics)
- Link / no link = complete / missing data

Missed matches can lead to...

Potential misclassification

Potential selection bias / loss of power

False matches can lead to measurement error or misclassification
Impact of missed links

- Missing data
- Misclassification or measurement error
- Erroneous inclusion/exclusion in an analysis
- ‘Splitting’ of one person’s records into many
<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>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><strong>Stillbirth</strong></td>
<td>0.5</td>
<td>4.6</td>
<td>3.2</td>
</tr>
<tr>
<td><strong>Baby factors</strong></td>
<td></td>
<td></td>
<td></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</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>

Impact of missed links

- Missing data
- Misclassification and measurement error
- Erroneous inclusion/exclusion in an analysis
- ‘Splitting’ of one person’s records into many
Impact of missed links

- Missing data
- Misclassification and measurement error
- Erroneous inclusion/exclusion in an analysis
- ‘Splitting’ of one person’s records into many
Impact of false links

- Misclassification or measurement error
- Erroneous inclusion/exclusion in an analysis
- ‘Merging’ of multiple people’s records into one
### 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

**Highly sensitive**

**Highly specific**
Solutions: evaluating linkage error

Gold standard data
- Positive / negative controls
- Comparisons with external references

Comparisons of linked / unlinked records
- Or of high / low quality records

Quality control checks
- Implausible scenarios
Positive / negative controls

• Linking mortality records for prisoners known to still be alive (-)

• Linking birth registrations for pregnancies known to have an abortive outcome (-)

• Linking infection surveillance records for neonates with a clinical recording of infection in their admission record (+)
Comparisons with external reference data

Solutions: evaluating linkage error

- **Gold standard data**
  - Positive / negative controls
  - Comparisons with external references

- **Comparisons of linked / unlinked records**
  - Or of high / low quality records

- **Quality control checks**
  - Implausible scenarios
### High / low quality records

<table>
<thead>
<tr>
<th>Age group in years</th>
<th>Available and valid</th>
<th>Not available or invalid</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All N</td>
<td>%</td>
<td>All N</td>
</tr>
<tr>
<td>All</td>
<td>7538</td>
<td>1759</td>
<td>23.3</td>
</tr>
<tr>
<td>0 to 14</td>
<td>122</td>
<td>40</td>
<td>32.8</td>
</tr>
<tr>
<td>15 to 44</td>
<td>4724</td>
<td>990</td>
<td>21.0</td>
</tr>
<tr>
<td>45 to 64</td>
<td>1576</td>
<td>409</td>
<td>26.0</td>
</tr>
<tr>
<td>65 and over</td>
<td>1061</td>
<td>320</td>
<td>30.2</td>
</tr>
<tr>
<td>Missing**</td>
<td>55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sex of case</td>
<td>Available and valid</td>
<td>Not available or invalid</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2941</td>
<td>726</td>
<td>24.7</td>
</tr>
<tr>
<td>Male</td>
<td>4355</td>
<td>1012</td>
<td>23.2</td>
</tr>
<tr>
<td>Missing</td>
<td>242</td>
<td>21</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Solutions: evaluating linkage error

- **Gold standard data**
  - Positive / negative controls
  - Comparisons with external references

- **Comparisons of linked / unlinked records**
  - Or of high / low quality records

- **Quality control checks**
  - Implausible scenarios
## Quality control checks

- Use evidence that two records do not belong to the same person to identify false-matches
- Admission following death
- Linkage of prostate cancer records with female hospital records

<table>
<thead>
<tr>
<th></th>
<th>Not (n = 773,446)</th>
<th>Simultaneous Admission (N = 324)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>51.7%</td>
<td>56.8%</td>
<td>.07</td>
</tr>
<tr>
<td>Preterm</td>
<td>7.9%</td>
<td>15.1%</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>White</td>
<td>75.8%</td>
<td>66.8%</td>
<td>(ref)</td>
</tr>
<tr>
<td>Mixed</td>
<td>4.6%</td>
<td>6.0%</td>
<td>.09</td>
</tr>
<tr>
<td>Asian</td>
<td>11.1%</td>
<td>18.4%</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Black</td>
<td>5.3%</td>
<td>4.4%</td>
<td>.83</td>
</tr>
<tr>
<td>Chinese</td>
<td>0.6%</td>
<td>1.0%</td>
<td>.26</td>
</tr>
<tr>
<td>Other</td>
<td>2.7%</td>
<td>3.5%</td>
<td>.22</td>
</tr>
<tr>
<td>Multiple birth</td>
<td>3.5%</td>
<td>3.8%</td>
<td>.75</td>
</tr>
</tbody>
</table>

Summary

• Linkage with administrative data is extremely valuable and can be more efficient than traditional follow-up
  – Cohorts created entirely from linked administrative data can provide new resources on a much larger scale than previously possible

• Data quality and linkage errors can challenge the reliability of linked data for analysis
  – Probabilistic linkage methods can provide measures of certainty
  – Mechanisms for linkage errors can be complex

• Methods for handling linkage errors can lead to more robust research
  – Quantitative bias analysis
Acknowledgements

Harvey Goldstein, Ruth Gilbert, Jan van der Meulen, James Doidge, Angie Wade, Gareth Hagger-Johnson

Funding:

Wellcome Trust grant numbers 103975/Z/14/Z and 212953/Z/18/Z.