DATA LINKAGE WORKFLOWS

John ‘Mac’ McDonald
Centre for Longitudinal Studies
Institute of Education
Data linkage (also known as record linkage)

- for organising ONE dataset
  - data cleaning
  - removing duplicates

- for merging TWO OR MORE datasets
  - merging individual-level datasets
  - adding census data to survey data

- for master data management
  - linking new transactions/events to master records
## Identification of Duplicates Given Name, Address, Age

### Matching Information

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>John A Smith</td>
<td>16 Main Street</td>
<td>16</td>
</tr>
<tr>
<td>J H Smith</td>
<td>16 Main St</td>
<td>17</td>
</tr>
<tr>
<td>Javier Martinez</td>
<td>49 E Applecross Road</td>
<td>33</td>
</tr>
<tr>
<td>Haveir Marteeneez</td>
<td>49 Aplecross Raod</td>
<td>36</td>
</tr>
<tr>
<td>Gillian Jones</td>
<td>645 Reading Aev</td>
<td>22</td>
</tr>
<tr>
<td>Jilliam Brown</td>
<td>123 Norcross Blvd</td>
<td>43</td>
</tr>
</tbody>
</table>
Record linkage . . .

“[is] a solution to the problem of recognizing those records in two files which represent identical persons, objects, or events (said to be matched).”

Problem of record linkage

problem - **quickly and accurately** determining if pairs of records describe the same entity, but unique IDs to bring together the matching records are lacking

records must contain some common identifying information (**keys or matching variables**)

- unique identifier (ideal in theory)
- name and/or address
- age (DOB) and sex

N.B. for very large databases, processing time and accuracy are concerns and blocking can be used to reduce the total number of record pairs compared
Files A & B, record a in A & record b in B

<table>
<thead>
<tr>
<th>File A</th>
<th>File B</th>
</tr>
</thead>
<tbody>
<tr>
<td>matching variables</td>
<td>matching variables</td>
</tr>
<tr>
<td>(v_1) (\ldots) (v_K) (X)</td>
<td>(Y) (w_1) (\ldots) (w_K)</td>
</tr>
<tr>
<td>a</td>
<td>b</td>
</tr>
</tbody>
</table>
Phases of record linkage

1. Input DB A
2. Input DB B
3. Search space reduction A x B
4. Search Space C ⊆ A x B
5. Decision Model Application

Output:
- Match
- Possible - match
- Non-match
Data linkage

- **Data linkage is context specific**, e.g. US addresses, scientific bibliography entries

- one-to-one linkage vs one-to-many linkage vs many-to-many linkage

- No universal best method for data linkage

- Linkage protocol used should be documented
Linkage projects typically have three phases

- **pre-linkage**
  - data cleaning
  - processing data fields to recognize similarity

- **linkage phase: deciding whether two records are a**
  - duplicate
  - match (link)

- **post-linkage**
  - manual/clerical review of unlinked records
  - research using the linked data
Data linkage is a challenging problem because of

- errors, variations and missing data on the information used to link records
- differences in data captured and maintained by different databases, e.g. age versus DOB
- data dynamics and database (DB) dynamics as data regularly and routinely change over time
  - name changes due to marriage & divorce
  - address changes
Methodology of record linkage

- two distinct methodologies for data linkage
  - deterministic linkage methods involve exact one-to-one character matching of linkage variable(s)
  - probabilistic linkage methods involve the calculation of linkage likelihood or linkage weights estimated given all the observed agreements and disagreements of the data values of the linkage variable(s)
  - probabilistic linkage methods can lead to much better linkage than simple deterministic linkage methods
Methodology of record linkage . . .

- methods from computer science, statistics and operations research
- methods primarily implemented by computer scientists
- general purpose versus domain specific, e.g. US addresses, scientific bibliography entries
- software for standardizing and parsing names and addresses that are used in the matching identifies
  - standardizing: replacing words with consistent abbreviations, e.g. street = ST
  - parsing: decomposing a string into a set of string components which are individually compared
Data problems

- typos/mispelling
- letters or words out of order
- fused or split words
- missing or extra letters
- incomplete words
- extraneous information
- incorrect or missing punctuation
- abbreviations
- multiple errors
Methodology of record linkage . . .

- deterministic algorithms - unique key(s)
- probabilistic algorithms - model based
- data mining techniques, e.g. neural networks
- Bayesian methods
- fuzzy methods, e.g. search engine/wild cards
- Boolean or other rule based methods
- linguistic rules (names from different cultures)
- combination of algorithms
Deterministic linkage

- **simplest method of matching** - sort/merge
- exact matching ONLY works well if the linking data are perfect and present in all the databases you want to link
- works best when there is a single unique identifier (key)
- otherwise, matching based on sets of identifiers predetermined by the researcher
- identifiers have equal weight
- identifiers chosen by researcher or by availability
- **works best with high quality data, but yields less success than probabilistic linkage**
Deterministic linkage . . .

- deterministic matching links records
  - using a fixed set of matching variables
  - exact one-to-one character matching of linking variables

- sometimes only the first few characters of a field are used with a wildcard substituted for later characters
  - primitive, but widely implemented, approach to tolerating errors
  - Martin versus Martinez
Deterministic linkage . . .

- brings together record pairs very efficiently by sorting both files using common identifier(s), which is the idea of a key

- keys associated with concepts of sorting/indexing
- example keys: surname, first name and DOB

- problem
  - offer no unique, known and accurate ID
  - missing values and partial agreements are common
<table>
<thead>
<tr>
<th>Surname</th>
<th>Name</th>
<th>Day of B</th>
<th>Year of B</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>414138</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5321</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>14004</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>168</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3090</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>102</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>969</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>513</td>
</tr>
</tbody>
</table>
Probabilistic linkage

- each matching variable is compared and assigned a score (weight) based on how well it matches
- frequency analysis of data values is important
- uncommon value agreement stronger evidence for linkage, e.g. Rumplestilskin versus Smith
- calculates a score for each field that indicates, for any pair of records, how likely it is that they both refer to the same entity
- sum the scores over fields
- sort record pairs in order of their scores (weights)
Probabilistic linkage . . .

- cut off values for scores (weights) are used to distinguish between matches and non-matches.

- above a certain threshold, everything is a match (link)

- below a certain threshold, nothing is a match (nonmatch or nonlink)

- in between (grey area), possible match needs manual/clerical review
Probabilistic linkage . . .

- **total score for a link between any two records is the sum of the scores generated from matching individual fields**

- score assigned to a matching of individual fields
  - is based on the probability that a matching variable agrees given that a comparison pair is a match
  - M-probability - similar to “sensitivity”, i.e. the proportion of actual positives which are correctly identified
Probabilistic linkage . . .

- score assigned to a matching of individual fields
  - reduced by the probability that a matching variable agrees given that a comparison pair is not a match \((U = \text{unmatched})\)
  - \(U\)-probability - similar to “specificity”, i.e. the proportion of negatives which are correctly identified

- agreement argues for linkage

- disagreement argues against linkage

- full agreement stronger evidence for linkage than partial agreement
Probabilistic linkage ...  

- based on the probabilities of agreement or disagreement between the identifiers

- all identifiers do not have equal weight

- accurate linkage is mainly dependent on the amount of discriminating power inherent in the variables common to the records that need to be matched and ‘good’ data

Fellegi-Sunter model

- **true matches**
- **true non-matches**
- **false matches**
- **false non-matches**

- **no-decision region** (hold for human review)

- designate as definite non-match
- designate as definite match

- *true matches*
  - ○ true non-matches

- `sim(a, b)`
Density function

Increasing value of $r = \frac{m(\gamma)}{u(\gamma)}$

Error $\lambda$

$T_u$

$T_m$

Error $\mu$

$U^*$

$M^*$
REL AIS

- REL AIS (Record Linkage At IStat) toolkit
- an open source toolkit for building record linkage workflows
- JAVA based
- statistical methods implemented in R

http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/
Figure 1: The record linkage complexity

**RECORD LINKAGE**

**COMPARISON FUNCTION CHOICE**
- Exact
- Edit distance
- Smith-Waterman
- Q-grams
- Jaro string comparator
- Soundex code
- TF-IDF

**SEARCH SPACE REDUCTION**
- Sorted Neighbourhood Method
- Blocking
- Hierarchical Grouping

**PRE-PROCESSING**
- Conversion of upper/lower cases
- Replacement of null strings
- Standardization
- Parsing

**DECISION MODEL CHOICE**
- Fellegi & Sunter
- Deterministic
- Bayesian
- Knowledge – based
- Mixed
Figure 2: Examples of record linkage workflows
<table>
<thead>
<tr>
<th>Requirement</th>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data requirement</td>
<td>Hierarchical structure</td>
</tr>
<tr>
<td></td>
<td>Workflow iteration:</td>
</tr>
<tr>
<td></td>
<td>• Higher level (household)</td>
</tr>
<tr>
<td></td>
<td>• Lower level (person)</td>
</tr>
<tr>
<td></td>
<td>High quality Equality comparison function on most of the phases</td>
</tr>
<tr>
<td></td>
<td>Large data set Blocking and phase iteration</td>
</tr>
<tr>
<td>Application requirement</td>
<td>Not significant errors in matching process</td>
</tr>
<tr>
<td></td>
<td>Probabilistic model and clerical review phase</td>
</tr>
</tbody>
</table>

Figure 7: An example of a pattern for building record linkage workflows
Results using deterministic approach

1°Merge : (1,1,1,1) + on the 1°Merge-residuals 2°Merge : (1,0,1,1) + on the residuals 3°Merge : (1,1,0,1)

| X       | P(X=1|M) | P(X=1|U) |
|---------|---------|--------|
| Surname | 0.9853  | 0.0023 |
| Name    | 0.9650  | 0.0074 |
| Day of birth | 0.9825  | 0.0327 |
| Year of birth | 0.9889  | 0.0127 |

Observed FMR=0.005
Observed FNMR=0.06

True Linkage Status

<table>
<thead>
<tr>
<th></th>
<th>Matched</th>
<th>Not Matched</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched</td>
<td>538</td>
<td>3</td>
<td>541</td>
</tr>
<tr>
<td>Not Matched</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>573</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The linkage results are "appreciable" but the linkage errors are not well estimated.

Observed FMR=0.017 vs the expected 0.001

Observed FNMR=0.010 vs the expected 0.0001
What should the linker do to help the analyst?

What should the analyst know about the linkage and how should that information be used?

In our opinion it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly. Obviously the quality of the linkage effort may directly impact on any analysis done.

Scheuren F & Winkler W E (1993) Regression analysis of data files that are computer matched - part 1. Survey Methodology 19, 39-58
What should the matching variable(s) be?


- IDs are subject to problems of survey item non-response and measurement error
- 5 linkages: respondent-supplied NINO & 4 linkages using different combinations of sex, name, address and DOB
- as many linkages were made using non-NINO-based matches as were made using matches on NINO
- former were also relatively accurate when assessed in terms of false-positive and false-negative linkage rates
So, it's agreed—we go ahead with the information-matching.