Linking data records using probabilistic techniques
Overview

• The choice between deterministic and probabilistic linkage methods

• Demonstration of probabilistic linkage software -- Links

• Conclusions regarding probabilistic and deterministic methods
The Massachusetts Maternal Mortality and Morbidity Project
Deterministic Linkage Challenges

• Which are the best variables to link with?
• What is an objective way to decide matched v. unmatched?
• When do we say “enough is enough”? 
The MA Pregnancy to Early Life Longitudinal Project (PELL)

Administrative Data

- Birth Certificate/Fetal Death
- Hospital Discharge
- Observational Stay
- Emergency Department
The MA Pregnancy to Early Life Longitudinal Project (PELL)

Programmatic Data

- MA Healthy Start
- Early Intervention
- WIC

Birth Certificate/Fetal Death
The MA Pregnancy to Early Life Longitudinal Project (PELL)

Vital and Health Statistics Data

- Birth Certificate/Fetal Death
- Death Data
- Birth Defects
PELL Data Set

- Used for many different analyses
  - Program Review
  - Surveillance
  - Research

How do we create a linked data set with flexibility?
New Challenges

- Reduce amount of time spent on each linkage
- Use one linkage algorithm for multiple years of data
- Deal with matched and unmatched in a consistent and objective way
- But *how*?
Probabilistic Record Linkage

- Uses probabilities to determine whether a pair of records refer to the same individual
- Calculates weights to quantify the likelihood that a pair of records are a true match
- Probabilistic weights may be either non-specific or value specific
General (Non-Specific) Weights

- Agreement on a specific variable
- Example:
  - **Agreement** on date of birth receives a higher weight than match on sex
  - **Disagreement** on sex receives a higher penalty than disagreement on date of birth
Value Specific Weights

- Agreement on a specific value of the variable being compared
- Example: Comparing initials using value specific weights
  - Agreement on initial Z receives higher weight than match on initial S
  - Disagreement on initial S receives higher penalty than disagreement on Z
Benefit of Weights

- Weights objectively reflect our confidence in a match
- Individual choice in cutting off low weights
Probabilistic Linkage Methods

- Some SAS programmers write their own probabilistic code
- Software packages
  - Very expensive
  - Difficult to use
  - Some applications are available as freeware or shareware
Choosing Probabilistic Software

<table>
<thead>
<tr>
<th>OS</th>
<th>Initial ($)</th>
<th>Yearly ($)</th>
<th>Link Type</th>
<th>Description</th>
<th>Audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatch (Integrity)</td>
<td>$100,000</td>
<td>???</td>
<td>Probabilistic</td>
<td>GUI</td>
<td>Marketing</td>
</tr>
<tr>
<td>Windows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generalized Record Linkage System (GRLS)</td>
<td>$18,800</td>
<td>10%</td>
<td>Probabilistic</td>
<td>ORACLE</td>
<td>Health care</td>
</tr>
<tr>
<td>UNIX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LinkPro</td>
<td></td>
<td>None</td>
<td>Determ &amp; Prob</td>
<td>SAS</td>
<td>Health care</td>
</tr>
<tr>
<td>Windows/Mainframe</td>
<td>$1,455 / $1,190</td>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Links: same as LinkPro but freeware
- FEBRL: also freeware, opensource
LinkPro Features

- Inexpensive
- Easy to use
- Created for health care record linkage
- Both deterministic and probabilistic linkage
LinkPro Features

- Capacity to recognize and accommodate duplicate records
- Supports full and partial/conditional comparisons
LinkPro Features

- Runs on any mainframe, mini, workstation or PC with SAS 6.06 or higher
- Free technical support
How LinkPro Works

- Automatically calculates and applies non-specific probabilistic weights
- Weights estimate the likelihood that a pair of records from separate files correspond to the same individual
How Weights Are Calculated

Computed building on $\log_2$ of the odds or frequency ratio calculated for each variable

\[
\text{weight} = \log_2 \frac{\text{OUTCOME freq in LINKED pairs}}{\text{OUTCOME freq in UNLINKABLE pairs}}
\]
LinkPro Statements and Syntax

_LINKPRO DATA1= SAS-data-set
DATA2= SAS-data-set <options> ;
_VAR variable-list;
_BY variable-list;
_RUN;
_LINKPRO <options>

_LINKPRO DATA1= SAS-data-set <options>
_DATA2= SAS-data-set <options> ;
_VAR variable-list;
_BY variable-list;
_RUN;
_LINKPRO <options>

MIN = number

- Minimum number of variables that must agree in _VAR statement
_LINKPRO <options>

**USEMEM**
- Stores input data in memory for faster execution

**SIMPLE**
- Deterministic linkage only
PAIRS = SAS-data-set

- Creates data set containing all ‘linkable’ pairs (potential links)

RESOLVE = 1xN or Nx1

- Allows one to many match
_LINKPRO <options>

DEBUG

- Prints all SAS statements and messages for problem diagnosing
**_VAR statement**

```sas
_LINKPRO DATA1= SAS-data-set
DATA2= SAS-data-set <options> ;
_VAR variable-list;
_BY variable-list;
_RUN;
```
_VAR statement

- Lists all variables used in linkage
- Numeric or character variables
VAR statement

Partial/Conditional Comparisons

- VAR1|VAR2
- VAR1 compared for agreement, if no match, VAR2 compared
- 3 possible outcomes and weights
_BY statement

_LIBKPRO DATA1= SAS-data-set
DATA2= SAS-data-set <options> ;
_VAR variable-list;
 BY variable-list;
_RUN;
_BY statement

- Optional statement
- Variable(s) that must match exactly
- Speeds up linkage
_LPX and _WTX statements

_LINKPRO DATA1= SAS-data-set
DATA2= SAS-data-set <options> ;
_VAR variable-list;
_BY variable-list;
_LPX 'SAS-statement(s)';
_WTX 'SAS-statement(s)';
_RUN;
Optional statement

Inserts SAS statements into the data step that generates linkable pairs

```
_LPX 'if given1^=given2 and given1>given2 then
_matched = _matched+1;';
```
• Optional statement

• Inserts SAS statements into the data step that calculates probabilistic weights

```sas
_WTX 'if abs(birthyr1-birthyr2)<=3 then
    _wgt=_wgt+1;';
```
LinkPro Output Files

1. LKD - Linked records
2. TIE - Links that could not be resolved
3. DAT1 - Unlinked records from the first data set
4. DAT2 - Unlinked records from the second data set
LinkPro Versus Deterministic

- Replicated original BC-HD link using LinkPro
- Found over 99% agreement in resulting links from two linkage methods
Benefits Of Probabilistic Linkage

- Routinized linkage process
- Provided objective way to deal with matched and unmatched data
- Reduced amount of code and time spent on linking data
- Able to inspect tied records or print out those with lowest 5-10% weight
Abandon Deterministic Altogether?

Definitely NOT

Choice of deterministic or probabilistic methods depends on:
- Type of project
- Data
Conclusions

- Deterministic vs. Probabilistic
  - Depends on your situation and goals
- Probabilistic linkage software can be affordable and easy to use!
More Information on LinkPro

http://members.shaw.ca/andre.wajda/linkpro.html

André Wajda
Andre.Wajda@shaw.ca
More Information on Links

Randy Walld

Randy_Walld@cpe.umanitoba.ca