Data Linkage: Improving, Integrating, and Maximizing Public Health Data

Texas Cancer Registry
Cancer Epidemiology and Surveillance Branch
Texas Department of State Health Services

Melanie Williams, Ph.D.
March 25, 2011
Background

• One of the primary functions of a cancer registry is to bring information together describing the same individual from a variety of data sources.

• This primary function applies to many aspects of public health as well. Multiple sources of information are still needed, health conditions are complex, and often intertwined.

• No data collection system is perfect in either quantity, quality, sources, or types of information we collect.
Background

• More stress is being placed on using high quality complex, integrated data to guide our public health policy, prevention, control, investigation, and research efforts.

• We are continually being asked to do more with less in supporting data collection efforts and customer requests.

• We must remain cognizant of the reporting burden we place on data reporters/stakeholders.

• We must safeguard the confidentiality of individuals to the fullest extent possible.

• Data linkage is an extremely helpful tool to help address these issues.
What do we often mean by the term “data linkage?”

• Bringing together two or more sources of data to compare and often improve common data elements, and/or add records or certain data elements contained in one file but not the other.

• Common applications include de-duplication, geographic coding, identifying missed cases and information, patient follow-up, and epidemiologic research.
Data Linkage

• Record linkage is a fundamental activity for cancer registries
  – Casefinding, finding other missing/unknown data, duplicate detection, follow-up, special studies

• Failure in the linkage process leads to
  – Over- or under-counting of cancers
  – Generation of inaccurate counts and rates
  – Incomplete information (e.g., vital status)
Definition

What is “data linkage?”

• First extensively developed at the U.S Census Bureau.

• By matching individuals counted in a census to those counted in an independent post or pre-enumeration survey, estimates of the quality of the enumeration could be produced.

Definition

Two main types of data linkage:

• External Linkage
  – Link one file to another file

• Deduplication
  – Special case of record linkage
  – Records in the same file are blocked, compared, and scored against each other
  – Results are ranked in list of record pairs
  – High-scoring pairs may be duplicates
Data Linkage Methods

• Various methods of data linkage
  – Ad hoc
  – Manual
  – Computerized Deterministic
  – Computerized Probabalistic
Data Linkage Methods

• Data errors are a fact of life
  – Missing SSN and other identifiers
  – Misspelled names and typo errors
  – Swap of birth day and birth month
  – Inconsistent ways of recording data (example: middle initial versus full name, ethnic surnames)
Data Linkage Methods

• Ad hoc methods abound despite inconsistencies and errors.

• Humans can conduct record linkage manually by visually comparing records from two separate sources.

• Approach is expensive, time consuming, tedious, and often impossible.
  – Large volume of data (often millions)
  – Rapid changes of data (updating existing records and adding new records)
Data Linkage Methods

- Record linkage is becoming easier
- Efficiency is a key feature
  - Faster, more efficient linkage process allows more linkages for less $$ and staff time
    - More accurate counts
    - More utility
    - Increased utilization of data
Data Linkage Methods

Basic Record Linkage Requirements

- Record linkage methodology
- Functionalities (2 file linkage, 1 file, batch/interactive processing)
- Fit to hardware/software environment
- Readiness for use
- Cost and simplicity
- Maintenance and support
Data Linkage Methods

Other Considerations

• Special features such as standardization of names and addresses
• Interactive review of “possible” matches
• Report generation
• Extraction of linked data
Data Linkage Methods

Other Considerations

• We must ensure confidentiality requirements and any other data restrictions are met

• Other issues must sometimes also be considered
  – Program staff
  – IT
  – Program attorneys, contract management
  – The IRB
  – Internal/external customers
Data Linkage Methods

Fields commonly used:

- SSN
- Surname, maiden, first, and middle names
- DOB
- Sex
- Race
- Other fields can be used as well, such as address

Data quality and completeness of these fields matters!
Deterministic versus probabilistic methods of linkage

- Deterministic record linkage does not capture many possible matches
- Probabilistic record linkage standard among cancer registries
Data Linkage Methods

“Exact Match” – Deterministic Linkage

• Simpler method of matching

• Records agreeing exactly within an individual data field or a group of common fields between records

• Approach relies on files having unique identifying information
  – SSN, surnames, given names, medical record numbers
  – Minimal amount of missing or inaccurate information
Data Linkage Methods

“Exact Match” – Deterministic Linkage

• Primary advantages

  – Brings together record pairs simply by sorting both files using a common unique identifier as a key field

  – Can be successfully applied when accurate records with unique identifying information are available
Data Linkage Methods

“Exact Match” – Deterministic Linkage

• Primary disadvantages

  – Absence/incompleteness/inaccuracy of key identifiable variables (surnames, birthdates, SSNs)
  – Spelling transcription errors in the data
  – Use of nicknames and proper names, change of name over time
  – Specialized code for deterministic combinations often is expensive, takes years to develop, and does not quite ever fulfill its goals. Flexibility is also lost.
Data Linkage Methods

Probabilistic Linkage

• Estimates the probability/likelihood that two records are from the same person versus not

• Recommended over traditional deterministic methods when:
  – Coding errors, reporting variations, missing data, or duplicate records encountered in data

• Frequency analysis of data values involved (important)
Data Linkage Methods

Frequency Analysis Examples

• How common is the name “Takaharu” in the Texas Cancer Registry versus in the Tokyo Cancer Registry

• If you have an “iffy” match and the surname is “Rumplepinder,” the SSN is missing and the dob is wrong, will you take it?

• What if you have the same iffy match but the surname is “Williams?”
Data Linkage Methods

Probabilistic Linkage, cont’d.

- Agreement on an uncommon value argues more strongly for linkage than a common value (e.g., Takaharu versus Williams in Texas)

- Agreement on a more specific attribute argues more strongly for linkage than agreement on a less specific one (e.g., SSN vs sex)

- Agreement on more attributes, disagreement on few, supports linkage
Data Linkage Methods

De-Duplicating/Linking Process

• Input: Records from database(s)

• Blocking: Probabilistic step that reduces number of comparisons between files, finds possible links by excluding majority of unlikely links
  – Often use records for first comparison with same soundex, SSN, and DOB
Data Linkage Methods

Weighting: A critical step in probabilistic linkage

- Once comparisons within blocks are made weights are calculated for each field comparison and a total weight is derived.
- Defines thresholds for automatically accepting or rejecting a linkage.
  - Gray area/marginal linkages can be reviewed.
Data Linkage Methods

Three Types of Agreement

• Simple agreement/disagreement
  – Discriminatory power varies among variables. A pair of records having different SSNs but the same surname are less likely to be a duplication than a pair having different surnames but the same SSN.
Data Linkage Methods

Three Types of Agreement

• Value-specific agreement
  – Discriminatory power also varies among values in the same variable. A pair of records with agreement on a common surname “Williams” carries less discriminatory power than a pair with agreement on a rare surname like “Rumpultin.”
Data Linkage Methods

Three Types of Agreement

• Partial agreement (similarity)
  – A pair with first name = “Jess” and “Jesse” is more likely to be consistent of duplicate records than a pair with the first names “Jess” and “Thomas.”
Software Packages

- A variety of software are available.
- LinkPlus is a free probabilistic record linkage program developed by the National Program of Central Cancer Registries, CDC
- Combines ease of use and statistical sophistication
- Detects duplicates within a data file, or links two data files together
- Supports fixed width files, delimited files, and standard cancer registry files
- Provides support for manual review of uncertain matches
Link Plus

- Designed especially for cancer registry work
  - However, can be used with any data
- Mathematics largely hidden from user
- Practical default values supplied for many tasks
- Familiar Windows interface
- Includes Help and test examples
- Disadvantage: have had some file size issues in the past
## Link Plus Linkage Overview

### External Linkage Steps:

1. Select Data Type for File 1
2. Locate/Identify File 1
3. Data Import for File 1
4. Select Data Type for File 2
5. Locate/Identify File 2
6. Data Import for File 2
7. Select Blocking Variables & Phonetic System
8. Select Matching Variables & Matching Methods
9. Select ID Variables
10. Define Missing Values
11. Select Direct/EM Method
12. Enter Cut-off Value
13. Specify Linkage File Name and Location
14. Run Linkage
15. Perform Manual Review of Uncertain Matches
16. Export Merged File
Link Plus Linkage Configuration

Identify/Import Data Files

Specify Data Type: Fixed Width

File 1: C:\RegPlus\LinkPlus\data\STATEVS2005.dat
File 2: C:\RegPlus\LinkPlus\data\CCRQ10.dat

Select blocking variables

<table>
<thead>
<tr>
<th>Data Item (File 1)</th>
<th>Data Item (File 2)</th>
<th>Phonetic System</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOB</td>
<td>Birth Date</td>
<td>Soundex</td>
</tr>
<tr>
<td>LNAME</td>
<td>Name--Last</td>
<td></td>
</tr>
<tr>
<td>SSN</td>
<td>Social Security Number</td>
<td></td>
</tr>
</tbody>
</table>

Select ID variables (File 1)

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Phonetic System</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCERT</td>
<td></td>
</tr>
</tbody>
</table>

Select matching variables and methods

<table>
<thead>
<tr>
<th>Data Item (File 1)</th>
<th>Data Item (File 2)</th>
<th>Matching Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOB</td>
<td>Birth Date</td>
<td>Date</td>
</tr>
<tr>
<td>LNAME</td>
<td>Name--Last</td>
<td>Last Name</td>
</tr>
<tr>
<td>FNAME</td>
<td>Name--First</td>
<td>First Name</td>
</tr>
<tr>
<td>SSN</td>
<td>Social Security Number</td>
<td>SSN</td>
</tr>
<tr>
<td>MI</td>
<td>Name--Middle</td>
<td>Middle Name</td>
</tr>
<tr>
<td>RACE</td>
<td>Race 1</td>
<td>Exact</td>
</tr>
<tr>
<td>SEX</td>
<td>Sex</td>
<td>Exact</td>
</tr>
</tbody>
</table>

Specify Missing Values

<table>
<thead>
<tr>
<th>Missing Value (File 1)</th>
<th>Missing Value (File 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day</td>
<td>Day</td>
</tr>
<tr>
<td>Month</td>
<td>Month</td>
</tr>
<tr>
<td>Year</td>
<td>Year</td>
</tr>
<tr>
<td>Format</td>
<td>Format</td>
</tr>
</tbody>
</table>

Direct Method/EM Algorithm

Cutoff Value: 7
Enter Cutoff

Results will be saved to
C:\RegPlus\LinkPlus\Report\ExternalLink.txt

Specify Linkage File Name and Location

Save Linkage Configuration

Run Linkage!
| Score | Class | Link ID | File | Record # | LNAME;Name-Last | FNAME;Name-First | DOB;Birth Dat | SSN;Social Security Num | ML;Name-Middle | SEX;Sex | RACE;Race | DT |
|-------|-------|---------|------|-----------|-----------------|------------------|--------------|------------------------|----------------|--------|-----------|----|-------|
| 18.7  | 4     | 60      | 1    | 41        | JONES GINA      | 01071937         | 806126065     | C                      | 2              | 1      | 0         |    |       |
| 18.6  | 4     | 61      | 1    | 27        | FOSTER LINDA    | 12011928         | 836926266      | B                      |                |        |           |    |       |
| 16.4  | 4     | 62      | 1    | 40        | LONG NORMAN     | 11051933         | 801825875     | BARBARA                |                |        |           |    |       |
| 18.4  | 4     | 63      | 1    | 21        | RICKARD AUDREY  | 01141921         | 801624953     |                        |                |        |           |    |       |
| 10.2  | 4     | 64      | 1    | 91        | AVER DAVID      | 09151930         | 806228612     |                        |                |        |           |    |       |
| 18.2  | 4     | 65      | 1    | 89        | KEEL DAVID      | 08221922         | 822228864     | J                      |                |        |           |    |       |
| 18.2  | 4     | 66      | 1    | 72        | EDWARDS ALLISON | 04051931         | 911630798     |                        |                |        |           |    |       |
| 10.2  | 4     | 67      | 1    | 66        | CHAPPELL WILLIAM| 12041924         | 812127511     | A                      |                |        |           |    |       |
| 15.0  | 4     | 68      | 1    | 4        | COGGIN BARBARA  | 01271940         | 870224354     | COGGIN BARBARA         |                |        |           |    |       |
| 14.0  | 4     | 69      | 1    | 9        | PRICE CLARE     | 08161926         | 81214719      | PRICE CLARE            |                |        |           |    |       |
| 17.9  | 4     | 70      | 1    | 78        | SPICER CATHERINE| 06301927         | 802827835     |                        |                |        |           |    |       |
| 17.9  | 4     | 71      | 1    | 64        | MAYNARD JOHN    | 03221956         | 884129670     |                        |                |        |           |    |       |
| 17.9  | 4     | 72      | 1    | 39        | SENA JOHN       | 03101913         | 807425964     |                        |                |        |           |    |       |
| 17.9  | 4     | 73      | 1    | 32        | NORRIS MICHAEL  | 06931967         | 800025427     |                        |                |        |           |    |       |
| 17.9  | 4     | 74      | 1    | 16502     | NORRIS MICHAEL  | 06031957         | 800025427     |                        |                |        |           |    |       |
Linkage Example

Cancer Registry data for John Smith:

<table>
<thead>
<tr>
<th>Last name</th>
<th>First Name</th>
<th>Site</th>
<th>SSN</th>
<th>DOB</th>
<th>Sex</th>
<th>DateDx</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMITH</td>
<td>JOHN</td>
<td>C619</td>
<td>123654789</td>
<td>0211934</td>
<td>1</td>
<td>06152004</td>
</tr>
</tbody>
</table>

Vital Statistics data for John Smith:

<table>
<thead>
<tr>
<th>Last name</th>
<th>First Name</th>
<th>DOB</th>
<th>Date of Death</th>
<th>COD</th>
<th>Death Cert #</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMITH</td>
<td>JOHN</td>
<td>02011934</td>
<td>03202006</td>
<td>123654789</td>
<td>01234</td>
</tr>
</tbody>
</table>

Linked data for John Smith:

<table>
<thead>
<tr>
<th>Last name</th>
<th>First Name</th>
<th>Site</th>
<th>SSN</th>
<th>DOB</th>
<th>Sex</th>
<th>DateDx</th>
<th>Death Date</th>
<th>COD</th>
<th>Death Cert #</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMITH</td>
<td>JOHN</td>
<td>C619</td>
<td>123654789</td>
<td>02011934</td>
<td>1</td>
<td>06152004</td>
<td>03202006</td>
<td>C100</td>
<td>01234</td>
</tr>
</tbody>
</table>
For more information on Texas cancer data or to make a request, contact:

Texas Cancer Registry
Texas Department of State Health Services
1100 W. 49th Street
Austin, Texas 78756
Mailing address: PO Box 149347, Austin, Tx 78714-9347
(512) 458-7523 –or- (800) 252-8059

Visit us on the Web:
http://www.dshs.state.tx.us/tcr