Integrating BigMatch into Automated Registry Record Linkage Operations

2014 NAACCR Annual Conference
June 25, 2014
Jason Jacob, MS, Isaac Hands, MPH, David Rust, MS
Kentucky Cancer Registry
Overview

• Record Linkage Basics
• Record Linkage in a Cancer Registry
• What is BigMatch ?
• Comparing BigMatch to LinkPlus
• Building a Master Patient Index
• Lessons Learned
• Software Demo
Record Linkage Basics

**Record Linkage:** Comparing 2 lists of patient records to find out which patients are the same

Exact matches are easy.

Partial matches can be hard:

- Is SSN or MRN enough?
- Is First + Middle + Last Name + DOB enough?
- Can transposed or smudged characters be handled?
- What about maiden names or nicknames?
- Which records do we compare?
  
  \[ 10,000 \times 100,000 = 10^9 \]
- How can we resolve ambiguities?
Record Linkage Basics

Blocking
– Which records will be compared?

Matching
– Which fields within the record will be scored?

Scoring Algorithm
– How will the scores be calculated?

Score Cutoff Value
– What is the minimum score for a true match?

Manual Review
– How many pairs must be reviewed?
Record Linkage in a Cancer Registry

• Casefinding
  – epath
  – Hospital discharge lists
  – disease indexes
  – Nonreportable lists

• Data from other registries

• Death clearance

• Follow-up Sources: SSA, CMS, NDI, other TLAs that help with data submissions and follow-up

• Research studies
Record Linkage in a Cancer Registry

- Most cancer registries use LinkPlus and SAS for record linkages
- KCR needs a way to automate record linkages, with minimal human intervention
- LinkPlus requires user interaction at multiple steps, does not run on Linux
- KCR is investigating BigMatch for automated record linkages
What is BigMatch?

“…a record linkage tool that can perform high speed, large file matching. The program can be used to match a moderate-sized file against several large files, and it can be used for deduplicating a file. The files do not have to be sorted prior to running the program.”


• Developed at the U.S. Census Bureau, freely available by contacting the authors:
  william.e.yancey@census.gov or william.e.winkler@census.gov.
What is BigMatch?

- Command-line driven software, does not provide graphical user interface

- Input files and Configuration file can be difficult to create:
  
  6 1 1 0 1 1 0 100 100  
  6 6 2 3 3 4  
  6 6 6 6 6 6  
  st 81 2 81 2 1  
  tract 88 3 88 3 1  
  first 64 13 64 13 0 uo 0.90 0.10

- Very fast results for very large files, but requires tuning
Comparing BigMatch to LinkPlus

- Using published methodologies and data sets, we compared the results of a BigMatch de-duplication with LinkPlus
  
  (http://www.sciencedirect.com/science/article/pii/S1532046411001729)

<table>
<thead>
<tr>
<th>Software</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigMatch</td>
<td>0.96</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td>LinkPlus</td>
<td>0.94</td>
<td>0.83</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Blocking: first name initial and Soundex of NYSIIS code for last name, DOB
Matching: last name, first name, dob, sex, postal code

Precision = True Matches / (True Matches + False Positives)
Recall = True Matches / (True Matches + False Negatives)
F-Measure = 2* (Precision * Recall) / (Precision + Recall)
Comparing BigMatch to LinkPlus

• Using LinkPlus to identify true matches from a Death Certificate linkage, we observed very good results from BigMatch

<table>
<thead>
<tr>
<th>Software</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>BigMatch</td>
<td>0.993</td>
<td>0.995</td>
<td>0.994</td>
</tr>
</tbody>
</table>

Blocking: SSN, DOB, Soundex LastName
Matching: LastName, FirstName, MiddleName, DOB, Sex, SSN

Total Death Certificates: 70,000
Total Registry Records: 300,000
Comparing BigMatch to LinkPlus

<table>
<thead>
<tr>
<th></th>
<th>BigMatch</th>
<th></th>
<th>LinkPlus</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pro</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fast</td>
<td></td>
<td>Fast</td>
<td>Fast</td>
<td></td>
</tr>
<tr>
<td>Free</td>
<td></td>
<td>No Graphical Interface</td>
<td>Free</td>
<td></td>
</tr>
<tr>
<td>Good Accuracy</td>
<td></td>
<td>Hard to configure</td>
<td>Good Accuracy</td>
<td></td>
</tr>
<tr>
<td>Scriptable</td>
<td></td>
<td></td>
<td></td>
<td>Easier to configure</td>
</tr>
<tr>
<td>Multi-platform</td>
<td></td>
<td></td>
<td>Graphical Interface</td>
<td></td>
</tr>
</tbody>
</table>

| **Con**        |          |                      |          |                      |
|                | Fast     | Not Easily Scriptable |          |                      |
|                | Free     | Windows only         |          |                      |
|                |          |                      |          | Easier to configure  |
Building a Master Patient Index

• KCR has a continuous feed of 4 different message types from 99 reporting facilities:
  – ePath
  – Hospital Discharge Reports
  – HL7 ADT Discharge Messages
  – Oncology Appointment Records

• All messages are stored in a relational database
  – About 3 million total messages
  – 200-3000 new messages each day
  – Multiple messages for each patient
Building a Master Patient Index

• How many unique patients are represented by our 3 million messages?

• Each patient should have a unique ID

• As we receive new messages for the same patient, we want to assign the same patient ID
Building a Master Patient Index

1. Messages are posted to database daily
2. BigMatch de-duplication runs every night
3. New patients are given new IDs
4. Messages for existing patients are assigned existing MPI IDs
5. Ambiguous matches are set aside for manual review
Lessons Learned

• Creating a Master Patient Index is time consuming and requires constant refinement
• Need staff resources for manual review
• Need to develop graphical interface for BigMatch manual review that can feed back into database
• Configuring linkage parameters is difficult
• BigMatch can be fast – BigMatch can be slow
  – Depends on configuration parameters
Lessons Learned

Not everything can be automated.

✓ Input file creation
✓ Configuration file creation
✓ Running Linkage
✓ Process Definite Matches

X Handling Ambiguous Matches
What Next?

- Refine configuration and performance
- Refine interface for manual review
- Commit staff resources for manual review
- Identify other registry processes for automated record linkage
- Investigate other record linkage engines
  - Scriptable
  - Cross-Platform
Demo