Data linkages in PEDSnet

Toan C. Ong, PhD
Assistant Professor
Department of Pediatrics
University of Colorado, Anschutz Medical Campus
Content

• Record linkage background

• Express Scripts and WashU linkage exercise

• Colorado and CHOP linkage exercise

• D2V record linkage project
Definitions

• Record linkage: The process of linking records that represent the same entity in one or more databases

• Privacy-preserving record linkage (PPRL): record linkage without revealing clear-text linkage data using data encryption
Objectives

Dataset A: EHR data
- Patient 1
- Patient 2
- Patient 3
- Patient 4
- Patient 5
- Patient 6

Dataset B: Claims data
- Patient 1
- Patient 2
- Patient 5

De-duplication

Enrichment
PPRL process

Christen & Verykios, 2012
Linkage data

Linkage variables

- Individual variables: First name, Last name, DOB, SSN, Gender, Zip code
- Combinations:

1. Seeded HashID of (First Name + Last Name + Date of Birth),
2. Seeded HashID of (Date of Birth + SSN),
3. Seeded HashID of (Last Name + SSN), or
4. Seeded HashID of (Three Letter First Name + Three Letter Last Name + Soundex First Name + Soundex Last Name + Date of Birth + SSN).

Abel et al., 2015
Data pre-processing

• Data standardization
  • Reformat linkage values
  • Remove special characters
  • Perform phonetic encoding
  • Name and address standardization
  • Missing data standardization
  • ...

Data pre-processing

• Data encryption/hashing
  • Advanced encryption standard (AES) methods
  • One-way hashing methods
  • Bloom filters, locality-sensitivity hashing

\[
\text{Dice coefficient} = \frac{2|X \cap Y|}{|X| + |Y|}
\]
PPRL process

Christen & Verykios, 2012
Indexing/searching

• Blocking: partition records into “blocks” and only perform linkage within blocks
  + significantly reduce number of comparisons
  - increase the probability of false negative linkage

• Blocking values
  • Blocking value 1: Birth year
  • Blocking value 2: Zip code
  • Blocking value 3: 4 initial letter of last name + birth year
  • Blocking value 4: Zip code + Soundex of last name

• Blocking performance:
  • Recall: proportion of true matches that the blocking criteria preserves
  • Reduction ratio: proportion of record-pairs discarded by the blocking
Comparison/linkage

• Deterministic
Probabilistic

Dataset 1: EHR Data

Dataset 2: Medicaid Claims Data
Classification

- Probabilistic

\[
match\_score = \sum_{i=1}^{n} w_i \cdot d_i
\]

- \( w_i \): Weight of linkage variable \( i \) – Value Range: 0-1
- \( d \): Edit distance (e.g. Levenshtein distance) – Value Range: 0-100

Two classes
- \( \text{match\_score} \geq \text{threshold} \) ➞ Matches
- \( \text{match\_score} < \text{threshold} \) ➞ Non-Matches

Three classes
- \( \text{match\_score} \geq \text{high threshold} \) ➞ Matches
- \( \text{match\_score} < \text{low threshold} \) ➞ Non-Matches
- Low threshold < \( \text{match\_score} < \text{high threshold} \) ➞ Potential Matches
Evaluation

- Number of true matches
- Number of false matches and non-matches

\[
\text{Precision} = \frac{\text{Total number of correct pairs found}}{\text{Total number of pairs found}}
\]

\[
\text{Recall} = \frac{\text{Total number of correct pairs found}}{\text{Total number of correct pairs}}
\]

\[
f - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]
Verification

• Synthetic datasets
• Gold standard datasets
• Manual adjudication
Founding Institutions
PEDSnet Patients
N = 5.14 million
Express Scripts – WashU linkage exercise

• Test phases
  • Pilot test (small datasets)
  • Full test

• Separate IRB protocols were approved for both tests
Data

• Pilot test
  • St. Louis Children’s: 404 records
  • Express Scripts (ESI): 581,607 records
    • Use 3-digit zip and birth year as the criteria to create a subset of Express Scripts dataset

• Full test
  • St. Louis Children’s: 78,250 records
  • Express Scripts (ESI): 9.2M records
Data preparation

• Standardization
  • Remove prefixes (e.g., Mr. Ms.) and suffixes (III, Jr).
  • Remove space, number and special characters from the names
  • Capitalize all text values

• Encryption
  • Use the CU record linkage (CURL) tool to encrypt clear-text data using the Bloom filter.
Process

1. Perform a **deterministic** linkage using the 5 variables and use the result as the baseline performance

2. Perform privacy preserving probabilistic record linkage on the encrypted datasets.

3. Examine record linkage performance at different acceptance scores and using different blocking schemes

4. Verify linkage results using clear-text data
Results

• Pilot test dataset size
  • SLC: 404
  • ESI: 581,607

• # of deterministic linkages: 238
• # of probabilistic linkages: 255
Results

• Verification findings:
  • Typos in the value of the linkage variables
  • Nick name
  • Middle name
  • Maiden name included in last name (two-word names)
  • Prefixes and suffixes
Results

• Full test dataset size:
  • SLC: 78,250 records
  • ESI: 9.2M records

<table>
<thead>
<tr>
<th>Total number of records: 78,250</th>
<th>Scheme 1: 4 initial letters of LN + YOB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scheme 2: 2 initial number of zip code + MOB + DOB</td>
</tr>
<tr>
<td></td>
<td>Scheme 3: Soundex of LN + initial of FN + DOB</td>
</tr>
<tr>
<td>&gt;=0.85</td>
<td>&gt;=0.90</td>
</tr>
<tr>
<td>Scheme 1</td>
<td>39,137</td>
</tr>
<tr>
<td>Scheme 2</td>
<td>34,844</td>
</tr>
<tr>
<td>Scheme 3</td>
<td>33,090</td>
</tr>
<tr>
<td>Scheme 1</td>
<td>Scheme 2</td>
</tr>
<tr>
<td>Scheme 1</td>
<td>Scheme 3</td>
</tr>
<tr>
<td>Scheme 2</td>
<td>Scheme 3</td>
</tr>
<tr>
<td>Scheme 1</td>
<td>Scheme 2</td>
</tr>
</tbody>
</table>
Linkage verification

• Linkage sample
  • Match score = 100 (highest possible). Randomly evaluate 5% cases.
  • Match score < 100 AND Match score >= 95. Randomly evaluate 20% cases.
  • Match score < 95 AND Match score >= 90. Randomly evaluate 50% cases.
  • Match score < 90 AND Match score >= 85 (lowest possible). Evaluate 100% cases.

• Adjudicators
  • 3 Students
Colorado-CHOP linkage

CHCO
Clear-text data

<table>
<thead>
<tr>
<th>FN</th>
<th>LN</th>
<th>DOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Doe</td>
<td>12/25/2012</td>
</tr>
<tr>
<td>Mary</td>
<td>Jean</td>
<td>05/12/2011</td>
</tr>
</tbody>
</table>

Hashed data (bit-strings)

1010... 1110... 0101... 1111... 0011... 1101...

CHCO Re-Identification

<table>
<thead>
<tr>
<th>CHCO_Random_ID</th>
<th>PEDSnet_ID</th>
<th>Match Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHCO_0001</td>
<td>PN_5399</td>
<td>0.97</td>
</tr>
<tr>
<td>CHCO_0002</td>
<td>PN_9722</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Encrypted linkage

Linkage verification (how???)

CHOP
Clear-text data

<table>
<thead>
<tr>
<th>FN</th>
<th>LN</th>
<th>DOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Doe</td>
<td>12/25/2012</td>
</tr>
<tr>
<td>Mary</td>
<td>Jean</td>
<td>05/12/2011</td>
</tr>
</tbody>
</table>

Hashed data (bit-strings)

1010... 1110... 0101... 1111... 0011... 1101...

CHOP Re-Identification

<table>
<thead>
<tr>
<th>CHOP_Random_ID</th>
<th>PEDSnet_ID</th>
<th>Match Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOP_0001</td>
<td>PN_1234</td>
<td>0.79</td>
</tr>
<tr>
<td>CHOP_0002</td>
<td>PN_5399</td>
<td>0.97</td>
</tr>
</tbody>
</table>
D2V record linkage project

- Position UC Denver as an honest broker for local, regional and national record linkage activities
- Provide record linkage as a service
- Develop and implement flexible, effective, efficient, and secure record linkage methods
Summary

• Why record linkage is not a simple problem?
  • Large amount of data
  • Data error
  • Data security and privacy

• What are the challenges of record linkage?
  • Clear-text linkage variables (SSN, first and last name, DOB...) are HIPAA-protected information
  • Attack to decrypt hashed data
  • Lack of gold-standard linked data to test record linkage methods
  • Difficult to perform linkage verification
Acknowledgements

• Record linkage teams in
  • SAFTInet
  • PEDSnet
  • pSCANNER
  • D2V

• Funding:
  • AHRQ
  • PCORI
• Thank you. Questions?