Data Linkage Approach with Data Flux Data Management Studio

An User’s Perspective

Sutapa Mukhopadhyay, PhD, April 1 2020
Introduction

- Methods to consider for record linkage
  - cleaning
  - standardization
  - matching algorithms

- Record linkage methods can influence completeness and accuracy of data and measures
  - important for analytics for both public health surveillance and research studies.
Objectives

- Evaluate the accuracy of approaches for record linkage
  - Deterministic vs Fuzzy
  - Cleaning and standardization

- Quantify the influence of these methods on opioid and related prescribing measures, key variables of interest in our public health analyses.

- To develop strategies for data linkage using Controlled Substance Monitoring Database (CSMD) and death certificate data

Software used: SAS Data Flux data Management studio
Typical Use of SAS Data Flux

- Data Profiling
- Data Standardization
- Data De-Duplication
- Data Linkage from various data sources

Benefits: Less programming, user friendly, visual, point and click menus, preview results while doing the job, direct access to big data sources
Available Identifiers in Death Certificate Data

- Decedents data includes:
  - **First Name, Last Name, Date of birth**
    - Name needs to be cleaned and standardized
  - **Middle Name**
    - Can be useful for common first/last names, if available
  - **SSN** (standardized with or without dashes)
  - **Sex**
  - **Address**
    - Address needs to be cleaned and standardized and geocoded.
Available Patient Identifiers in the CSMD

- There is no unique patient identifier (e.g., SSN) for most records.

- Patient data includes:
  
  - **First Name, Last Name, Date of Birth**
    - Name needs to be cleaned and standardized
  
  - **Middle Name**
    - Can be useful for common first/last names, if available
    - High proportion missing
  
  - **Sex** (data quality a concern; multiple values of sex for same patient)
  
  - **Address**
    - Patients move and have many addresses (and many inaccuracies in addresses)
    - Address needs to be cleaned, standardized and geocoded
Record Linkage Approach

- Comprehensive name cleaning using SAS
- Deterministic matching approach using SQL
  - Exact DOB, any last name, and any first name
- Probabilistic / fuzzy matching algorithms with varying sensitivity using Data Flux
Identifiers Used for Data Linkage in Data Flux

- First first name
- All last names
- Date of birth

After cleaning separate multipart name

Implications of Inaccurate Patient Record matches in CSMD

- Matches may be incorrect within a matched record set (i.e., wrong patient linked)

- Correct matches may be missing from a matched record set (i.e., missing one or more patient records that have relevant prescription information)

- Both impact accuracy of prescription history in the CSMD
Methods Overview: SAS DataFlux vs Deterministic Matching

- SAS/SQL programming\(^a\)
- SAS programming for cleaning and standardization/parsing\(^a\) and SAS DataFlux for fuzzy matching
  - Varying sensitivity\(^b\) values for names and DOB
- SAS DataFlux alone
  - Varying sensitivity\(^b\) values for names and DOB

\(^a\)Extensive name cleaning and standardization using SAS programming.

\(^b\)Sensitivity values from DataFlux are based on a combination of standardization and fuzzy matching techniques (e.g., phonetics, transformations) using natural language processing.

### Missed Matches: Outcomes of Interest

<table>
<thead>
<tr>
<th>Matching algorithm</th>
<th>Matched overdose deaths (Total deaths # 5,483)</th>
<th>Additional overdose deaths matches from previous step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>4,572</td>
<td>NA</td>
</tr>
<tr>
<td>Method 2</td>
<td>4,684</td>
<td>112</td>
</tr>
<tr>
<td>Method 3</td>
<td>4,714</td>
<td>30</td>
</tr>
</tbody>
</table>

- Method 1: Clean and standardized names in SAS, exact match on name and DOB using SQL programming
- Method 2: Name cleaning and standardization in SAS, fuzzy matching in SAS DMS (name sensitivity 85, **DOB sensitivity 95**)
- Method 3: Name cleaning and standardization in SAS, fuzzy matching in SAS DMS (name sensitivity 85, **DOB sensitivity 85**)
Identifiers Used for False Positive (FP) Check

- **Address**: Both original and geocoded address
- **SSN**: SSN from death files
  - Customer number from CSMD file (very few records have true social security number)
Protocol for False Positive Check

- A match was confirmed if first and last name and DOB of a patient were identical between death and CSMD.

- A match was also confirmed despite some small unmatched names (like one vs two first or last name) if addresses were identical between death and CSMD.

- A match was also confirmed despite some small unmatched names (like one vs two first or last name) if SSN were identical between death and CSMD.

---

Nechuta S. et al., 2020
After Manual Review

- Not FP (able to confirm a correct match)
- Possible FP (i.e., unable to confirm true match status based on available information)
- True FP (confident an incorrect match)
## Potential False Positives

<table>
<thead>
<tr>
<th>Matching algorithm</th>
<th>Number of matches (patient record from CSMD)</th>
<th>Possible False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>19,088</td>
<td>344</td>
</tr>
<tr>
<td>Method 2</td>
<td>20,316</td>
<td>713</td>
</tr>
<tr>
<td>Method 3</td>
<td>20,754</td>
<td>1,059</td>
</tr>
</tbody>
</table>

- non-match records had been manually checked.

Nechuta S. et al., 2020
Summary

• Data cleaning and standardization is critical first step
  – Cleaning and Standardization are the multi-step processes

• Fuzzy matching using SAS DataFlux improved:
  – Number of correct matches to eligible decedents
  – Number of correct matches to eligible prescriptions of interest

• Using multi-step algorithms with validation and manual review are absolutely necessary for false positive check

Nechuta S. et al., 2020
SUDORS Linkage to IMED* and CSMD

Select SUDORS cases from provisional death file (death certificate)

Pull records from IMED based on:
- Name
- Date of Death

Use that file for linkage to CSMD data using SAS DataFlux

Variable used for linkage and false positive check:
- First Name
- Last Name
- Date of Birth
- Resident Address (original and geocoded)
- SSN

*Interim Medical Examiner Database
Results After Linkage

Total records from SUDORS 972

After Linkage with CSMD

SUDORS records linked 844

Delete false positives

Final SUDORS records 844

Patient records from CSMD 4,330

Final Patient records from CSMD 4,309

Final Prescriptions 58,588
Lessons Learned

- Data linkage method protocols should include
  - cleaning and standardization (name, DOB, address, SSN)
  - manual review for false positive check
OIA Data Linkage Products

- Published papers:
  - Overdose death and prescription history (*Nechuta S et al, 2018*) – linkage via SQL programming
  - Methods paper about record linkage approaches (*Nechuta S et al, 2020*) - linkage via Dataflux
  - Non-fatal drug overdose and prescription history (*Krishnaswami S et al, 2020*) - linkage via Dataflux
  - Injured worker’s and prescription history (*Durand Z et al, 2019, Durand Z et al, 2019*) - linkage via SQL programming
Thank you!

Questions?

Contact: Sutapa.Mukhopadhyay@tn.gov

Office of Informatics and Analytics
Tennessee Department of Health