Management of Hepatitis C cases in New Jersey Department of Health Surveillance System

June 8, 2022
Presentation overview

- Snapshot of New Jersey
- Burden of Hepatitis C in New Jersey
- New Jersey Communicable Disease Service surveillance system, CDRSS
- Hepatitis C and CDRSS
- Data Management: Hepatitis C Registry and PS2103
- Data Management in CDRSS
- Manual Deduplication using Link Plus
- Benefits and Limitations of Link Plus
- Manual Deduplication: Link Plus Process
- Computing Duplicates in Link Plus
- Link Plus: Matching Variables
- Alternative System for deduplication of cases
New Jersey located in the Mid-Atlantic region and borders two metropolitan cities; New York City and Philadelphia

- Most densely populated state
- Population ~9.0 million; 22% foreign-born
- Jurisdictions of 21 counties and 565 municipalities
- Trenton is the capital; located in Mercer county
- Most people live close to the metropolitan cities and the Jersey shore
Burden of Hepatitis C in New Jersey

- Hepatitis C Acute, Chronic and Perinatal are reportable
- About 9000 reports for Hepatitis C received annually
- > 6000 newly reported Acute and Chronic cases yearly; a slight decrease is observed
- New reports in multiple generations
- Highest rate of Hepatitis C (HCV) mainly in the most populated areas
Communicable Disease Reporting and Surveillance in New Jersey

- CDRSS, NJDOH official disease surveillance system with over 100 reportable and non-reportable diseases and work-related conditions
- Integrated electronic lab reporting (ELR) system that is patient-centric and web-based
- Has advanced tools for data analysis, visualization, mapping and exportable reports
- Has the capacity to manage outbreak investigations, schools and other facility surveillance
- Structured to accept positive and non-positive reports according to case definitions
- Industry-standard security features
Hepatitis C and CDRSS

- At least **90% of HCV labs are reported electronically** in CDRSS
- Some manual entry from paper laboratory reports
- HCV Investigators provide updates on HCV reports for residents in their jurisdiction in CDRSS
- Multiple person and case merges per year not automatically identified
- **2779** cases manually merged in CDRSS, in 2021
Data Management: Hepatitis C Registry (PS21-2103)

• As part of the grant requirement, overall data quality is essential to maintain NJDOH HCV registry and avoid duplication of cases

• Management of jurisdictional laboratory-based case cascade for hepatitis to monitor testing to cure efforts

• New Jersey CDRSS has the capacity to capture negative lab reports – thus increase the need for deduplication, as more laboratories increase HCV RNA reflex testing

• NJDOH runs the deduplication program quarterly to reduce incidence of duplicates
  • Manual process when demographics are not matched one to one
  • Takes about 2.5 hours a week
Data Management in CDRSS

- Managing a patient-centric system
  - ELR rules for incoming reports
  - Disease/Condition algorithms: Confirmed, Probable, Not a Case
  - New cases or reinfections?

- Duplicate Persons and Cases
  - ELR deduplication
  - Deduplication by system merge: One to one demographics
  - Manual deduplication: Link Plus Software

How do duplicates get into the system?

Data Entry Errors (99% of errors)
- Demographics Discrepancies:
  - Name: Elizabeth/Liz/Elisabeth, O’Sullivan/Osullivan
  - DOB: 12/1/1999 – 12/11/1999
Manual Deduplication: Using Link Plus Software

• Link Plus is a CDC software originally designed as cancer record matching and deduplication tool for cancer registries

• Computes probabilistic record linkage (matching) and has the capacity to detect duplicates using a choice of two phonetic coding systems (Soundex and NYSIIS)

• The program works for any number of records with a limit of 4.5 to 4.8 million records

• Provides Best Match options of all potential data matches or duplicates based on user defined threshold (cut off value)

• Link Plus is **FREE: $0.00**
Benefits and Limitations of Link Plus

Benefits
• Designed especially for cancer registry work but can be used with ANY DATA
• Probabilistic math and computation largely hidden from user
• System default values provided deduplication
• Software includes robust “Help Menu” and test examples
• Output results are clear and accessible to new or advanced users
• ♫ FREE FREE FREE FREE...FREE ♫

Limitations
• Creating external data files to be imported
• Default values may need to be changed multiple times
• Duplicates identified will require manual review and merging
Manual Deduplication: Link Plus Process

Link Plus is REALLY EASY to use!

Step 1: Select data file

Step 2: Select blocking variables and phonetic system

Step 3: Click run...and the software will do the rest!

DATA FILE → Blocking → Paired Probabilistic Matching → Scored Output → Manual review
Computing Duplicates in Link Plus

Data for Deduplication

Blocking

Matching and Scoring

High Score

Gray Area

Low Score
Link Plus: Matching Variables

Demographic data
- First name
- Last name
- Middle name
- Date of birth
- Social Security Number (*not collected in CDRSS*)

Other matching variables
- Zip code
- Generic ID
- Exact string data
- Value specific data
## Data File

X:\DISEASE INFORMATION\HEPATITIS C\DATA & STATS (BY YEAR)!Dedup\2021 to 2017 with

### Select Blocking Variables

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Phonetic System</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAST_NAME</td>
<td>Soundex</td>
</tr>
<tr>
<td>FIRST_NAME</td>
<td>Soundex</td>
</tr>
<tr>
<td>* DOB</td>
<td>Soundex</td>
</tr>
</tbody>
</table>

### Select ID Variables

<table>
<thead>
<tr>
<th></th>
<th>DISEASE</th>
<th>PATIENT_ID</th>
<th>CASE_ID</th>
<th>GENDER</th>
<th>COUNTY_1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Select Matching Variables and Methods

<table>
<thead>
<tr>
<th>Data Item</th>
<th>Matching Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAST_NAME</td>
<td>Last Name</td>
</tr>
<tr>
<td>FIRST_NAME</td>
<td>First Name</td>
</tr>
<tr>
<td>DOB</td>
<td>Date</td>
</tr>
</tbody>
</table>

### Missing Value

- **Direct Method**: Cutoff Value 20
- **Results will be saved to**: C:\RegPlus\LinkPlus\report\LinkageReport.txt

---

**File for dedup**

**Other variables in data file**

**Blocking and matching variables**

**Manual review (if needed)**

**Output data file**
Linking Process

Field for Blocking
soundex on LAST_NAME
soundex on FIRST_NAME
soundex on DOB

Matching Parameters

<table>
<thead>
<tr>
<th>Matching Field</th>
<th>m-prob</th>
<th>u-prob</th>
<th>agree</th>
<th>disagree</th>
<th>matching method</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOB</td>
<td>0.86000</td>
<td>0.00088</td>
<td>8.54696</td>
<td>-2.92987</td>
<td>date</td>
</tr>
<tr>
<td>LAST_NAME</td>
<td>0.97000</td>
<td>0.00055</td>
<td>5.10070</td>
<td>-2.92672</td>
<td>last name</td>
</tr>
<tr>
<td>FIRST_NAME</td>
<td>0.96000</td>
<td>0.00054</td>
<td>5.10070</td>
<td>-2.92672</td>
<td>first name</td>
</tr>
</tbody>
</table>

m-prob: The probability that a matching variable agrees given that the comparison pair being examined is a match
u-prob: The probability that a matching variable agrees given that comparison pair being examined as a non-match
agree: The agreement weight assigned for an agreement on a given matching variable
disagree: The disagreement weight assigned for a disagreement on a given matching variable

matching method: The method used for computing the weight on a given matching variable.

Match Report

"Sorted by score descending (higher score, higher likelihood of being a matched pair)"
Number of Matched Pairs = 1014

<table>
<thead>
<tr>
<th>Linkage ID</th>
<th>Set ID</th>
<th>Record #</th>
<th>DISEASE</th>
<th>PATIENT_ID</th>
<th>CASE_ID</th>
<th>DOB</th>
<th>LAST_NAME</th>
<th>FIRST_NAME</th>
<th>COUNTY_1</th>
<th>MUNICIPALITY_1</th>
<th>Class</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1013</td>
<td>12518</td>
<td>HEPATITIS C</td>
<td>2638663</td>
<td>1229587</td>
<td>12/1/1999</td>
<td>O'SULLIVAN</td>
<td>ELIZABETH</td>
<td>PASSAIC</td>
<td>PASSAIC CITY</td>
<td>7</td>
<td>27.1</td>
</tr>
<tr>
<td>1</td>
<td>1013</td>
<td>17802</td>
<td>HEPATITIS C - CHRONIC</td>
<td>3143025</td>
<td>1373103</td>
<td>12/11/1999</td>
<td>O' SULLIVAN</td>
<td>ELISABETH</td>
<td>PASSAIC</td>
<td>PASSAIC CITY</td>
<td>7</td>
<td>27.1</td>
</tr>
<tr>
<td>2</td>
<td>1012</td>
<td>13423</td>
<td>HEPATITIS C - CHRONIC</td>
<td>3634218</td>
<td>1311037</td>
<td>7/4/2002</td>
<td>PATEL</td>
<td>SUUKKA A</td>
<td>HUDSON</td>
<td>SEACAUS TOWN</td>
<td>8</td>
<td>26.6</td>
</tr>
<tr>
<td>2</td>
<td>1012</td>
<td>30293</td>
<td>HEPATITIS C - ACUTE</td>
<td>3639671</td>
<td>1700103</td>
<td>7/4/2002</td>
<td>PATEL</td>
<td>SITUKA</td>
<td>ESSEX</td>
<td>NEWARK CITY</td>
<td>8</td>
<td>26.6</td>
</tr>
<tr>
<td>3</td>
<td>1011</td>
<td>17502</td>
<td>HEPATITIS C - CHRONIC</td>
<td>735475</td>
<td>1410110</td>
<td>1/1/1961</td>
<td>SMITH</td>
<td>JOHN</td>
<td>HUDSON</td>
<td>JERSEY CITY</td>
<td>10</td>
<td>26.8</td>
</tr>
<tr>
<td>3</td>
<td>1011</td>
<td>38174</td>
<td>HEPATITIS C - CHRONIC</td>
<td>9900128</td>
<td>5917395</td>
<td>1/1/1961</td>
<td>SMITH</td>
<td>JOHN</td>
<td>HUDSON</td>
<td>JERSEY CITY</td>
<td>10</td>
<td>26.8</td>
</tr>
</tbody>
</table>
Alternative system – Match*Pro

• Match*Pro is a Java-based application developed by Information Management Services, Inc (IMS) for probabilistic record linkages.

• Also based on the Fellegi and Sunter model

• It has a user-friendly tabbed interface for the configuration of linkages and the manual review of uncertain matches

• Match*Pro can run on any fixed-width or delimited data file as Link Plus but can also support XML files

• Match*Pro is flexible and allows users to filter incoming data, specify blocking and matching methods, define unknown values, adjust weights, identify cutoff points, and more during the deduplication process

• **Match*Pro has 9 distinct blocking strategies**
**Match*Pro vs Link Plus?**

- **Match*Pro** is more advanced than **Link Plus** and can handle more parameters for a complicated probabilistic record linkage or deduplication.

- Users have access to **more tools** during the manual review process including a Query Editor.

- Probabilistic math and computation visible to user during deduplication process.

- **Match*Pro** has the option to filter input file before deduplication process.

- Also **FREE** but requires registration before download.

- Overall, **Link Plus** is **easier** to use for everyday work.
How to get Link Plus

Download from CDC National Program of Cancer Registries (NPCR)

https://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm#installation

• Follow installation directions
Questions?

THANK YOU
Bernice Carr, M.P.H., M.S.
Epidemiologist- Hepatitis C
Communicable Disease Service
New Jersey Department of Health
(609) 826.5964
Bernice.Carr@doh.nj.gov

Mojisola Ojo, M.P.H
Epidemiologist- Data Coordinator
Communicable Disease Service
New Jersey Department of Health
(609) 826.5964
Mojisola.Ojo@doh.nj.gov