IMPACT OF DE-IDENTIFICATION ON MASTER PATIENT INDEX AND DATA LINKAGES

August 2020
Kathy Hines, Senior Director of Partner Operations & Data Compliance
Scott Curley, Manager of Privacy & Compliance
Motivation for Change

- Rising external cybersecurity threats to healthcare data
- Internal risks of accidental or intentional data exposure.
- Specific to the APCD – Federal Law 42 CFR Part II
Outright removing PII would prevent CHIA and our external community of data users from connecting health care encounters across carriers and to other datasets.

CHIA set an objective to dramatically decrease the risk of exposure of collected PII while retaining the ability to connect data together.
CHIA’s Solution

1. **Software**
   - CHIA’s *File Secure* software is deployed to the site of data submission (insurance carriers and hospitals) that replaces key PII fields with *pseudonymized* equivalents.

2. **Internal Architecture**
   - CHIA never receives PII “in the clear” and the data is stored separately from the data warehouse and are not released to internal users or external data applicants.

3. **Submission Guide Updates**
   - CHIA stopped collection of certain fields.

4. **Master Patient Index**
   - One ID for each person regardless of insurance carrier with the ability to link to external data.
DE-IDENTIFICATION USING EXPERT DETERMINATION
### HIPAA De-Identification

#### Safe Harbor

**Pros**
- Easy to implement and maintain

**Cons**
- 18 data elements redacted or removed entirely
- More restrictive than statistical de-identification with respect to birth dates, service dates, and geographic data

#### Expert Determination

**Pros**
- Methodology tailored to data set in question
- Lower overall risk of re-identification

**Cons**
- No single method for implementation
- Routine reassessment
- More restrictive than Safe Harbor with respect to some individual claim lines

*Slide courtesy of ONPOINT Health Data*
OnPoint Worked with CHIA to Define Approach

- Established the variables to be considered for a formal re-identification risk analysis
  - Catalogued all **direct identifiers** and **quasi-identifiers**
- Determined acceptable risk levels
  - Minimum cell size, maximum risk, average risk
  - Assumed an adversarial environment where the recipients of the data have knowledge of quasi-identifying values for the individual
- Established annual re-assessments

* Slide courtesy of ONPOINT Health Data
Applied the Data Strategy

- The risk mitigation model was applied to multiple years of data (MA APCD data set years 2012 – 2017) to assess the risk stability over time and project a solution for the following year.

* Slide courtesy of ONPOINT Health Data
FILE SECURE
CHIA’s File Secure

- Data Submitters prepare files that include PII at their location
- *File Secure* replaces key fields with *pseudonymized* values (128 character length) while still at their location
  - Name
  - SSN
  - Full DOB (MMYYYYY are left in the clear for analytics)
- “In the clear” versions of Name, SSN, DOB never leave the data submitter’s site
CHIA’s File Secure

- Zip code processing
  - Flag if invalid zip code
  - Retain MA Zip codes only
  - Map MA Zip codes to mask small areas in MA APCD

- State code processing
  - Flag if invalid state
  - Retain only New England and New York state codes
  - Map MA Zip codes to mask small areas in MA APCD

- *File Secure* encrypts the file with NIST compliant encryption before data is sent to CHIA
SUBMISSION GUIDE
Submission Guide Changes – Data Removal

- Claims
  - First/Last names
  - Social Security numbers (SSNs)
  - Address information

- Eligibility
  - Street/City address information
  - Zip code limited to 5 digits
  - Race/Ethnicity indicators
  - Disability/Marital/Student/Family size indicators
  - Language (list abbreviated)
  - Date of Death
Insurance Carrier Submissions

APCD Submission Files (“in the clear”)

- Eligibility Removed:
  - Street/City
  - Marital Status
  - Race/Ethnicity
  - Employee Status
  - Student Status
  - Date of Death

- Medical Claims Removed: PII

- Product

- Provider

- Dental

- Rx Claims

CHIA File
Secure Software

- USPS Nickname Table
- NYSIIS - First and Last name
- HASH Function
- Remove known dummy values
- Re-map 1% of population (small ZIP codes)

Eligibility
Final

- HASH
  - First Name
  - Last Name
  - SSN
  - DOB

- Clear
  - State
  - Zip
  - Insurance ID
  - Org ID
  - Month / Year of Birth
  - Gender

Encrypted for
Transport to CHIA

CHIA Landing Zone

Insurance Carrier Site
CHIA creates a master patient index (MPI) using a probabilistic matching algorithm with *pseudonymized* identifiers. The ID connects all records that are very likely the same person and assigns them a key that is not based in any way on PII or any other attributes of a person’s data.

Example of what an APCD data user might have access to

- MPI – CHIA’s randomly generated unique ID for a person
- MM/YYYY of birth
- 5 digit ZIP code for largely populated ZIP codes

CHIA has deployed a service to connect external data to APCD or Case Mix using a combination of CHIA’s *File Secure* software and CHIA’s probabilistic matching engine
### CHIA Master Patient Index

**CHIA Landing Zone**

**Data Preparation**
- Filter Known Data Issues

**Data Load**
Records where:
- Insurance ID
- Org ID are the same are considered the same person.

The last 5 valid values of each input field are stored to capture name changes, people moving etc.

**CHIA APCD Algorithm**
- Probabilistic matching method using:
  - HASH Fname
  - HASH Lname
  - HASH SSN
  - HASH DOB
  - Zip
  - Gender
- Links records within and across carriers.

**CHIA Master Patient Index Hub (MEID)**

### CHIA MPI Org ID Insurance ID First Name Last Name DOB Gender SSN ZIP Code

<table>
<thead>
<tr>
<th>CHIA MPI</th>
<th>Org ID</th>
<th>Insurance ID</th>
<th>First Name</th>
<th>Last Name</th>
<th>DOB</th>
<th>Gender</th>
<th>SSN</th>
<th>ZIP Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>111111</td>
<td>30</td>
<td>BBY00002211</td>
<td>ABCD</td>
<td>QRSTUVWXYZ</td>
<td>POIUYT</td>
<td>F</td>
<td>HFHDSFH</td>
<td>02116</td>
</tr>
<tr>
<td>112233</td>
<td>22</td>
<td>HW00000122</td>
<td>GWD</td>
<td>DGEQDFFGFG</td>
<td>GFDIFF</td>
<td>M</td>
<td>FGDDBFG</td>
<td>02118</td>
</tr>
<tr>
<td>112233</td>
<td>30</td>
<td>BBY00034234</td>
<td>GW3D</td>
<td>DGFDGFFGFG</td>
<td>GFDIFF</td>
<td>M</td>
<td>FGDDBFG</td>
<td>01906</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13116</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>91925</td>
</tr>
</tbody>
</table>
CHIA MATCHING SERVICE
CHIA Matching Service (Master Data Management)

Customer File
- First Name
- Last Name
- SSN
- Gender
- DOB
- Zip Code
- Insurance ID
- Study ID

The more complete the file, the better the match results however not all fields are needed for each record for a confident match.

CHIA File
Secure Software

Customer Site

CHIA File
Prepped
- HASH
  - First Name
  - Last Name
  - SSN
  - DOB

Linking File
Prepped
- In the Clear
  - Zip
  - Insurance ID
  - Gender
  - Study ID

_used by customer_

CHIA APCD Algorithm

Scores each input record against likely candidates in the MPI Hub.

High Score Matches
- Custom Match Threshold based on how accurate the matches need to be.
  - For example: Higher = All fields present and up to 1 mismatch

Lower Score Matches
- Any number of additional match scenarios can be added and separated from the High Score Matches based on study need.
  - For example: Lower = SSN Missing and up to 1 mismatch

CHIA use Master Enterprise ID to identify corresponding claims, this ID is then replaced with the project’s unique Study ID and claims returned to customer.

SFTP to CHIA

CHIA

Impact of De-identification on MPI and Data Linkages
| Scott Curley, Kathy Hines | August 2020
**CHIA Linkage Service (MPI Search)**

Input Row from Customer - Hashed Equivalent

<table>
<thead>
<tr>
<th>Study ID</th>
<th>First Name</th>
<th>Last Name</th>
<th>DOB</th>
<th>SSN</th>
<th>Zip Code</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>8888</td>
<td>ABCD</td>
<td>QRSTUVWXYZ</td>
<td>POIUYT</td>
<td>02116</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>

**APCD Linking Scenarios**

<table>
<thead>
<tr>
<th>CHIA ID (MPI)</th>
<th>First Name</th>
<th>Last Name</th>
<th>DOB</th>
<th>SSN</th>
<th>Zip Code</th>
<th>Gender</th>
<th>Match Result</th>
<th>Match Score</th>
<th>Disposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>4455544</td>
<td>ABCD</td>
<td>QRSTUVWXYZ</td>
<td>POIUYT</td>
<td>02116</td>
<td>F</td>
<td>5 Matches, 0 Mismatch</td>
<td>Highest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4455544</td>
<td>ABCD</td>
<td>QRSTUVWXYZ</td>
<td>POIUYT</td>
<td>02119</td>
<td>F</td>
<td>4 Matches, 1 Mismatch</td>
<td>Higher</td>
<td>Input Row links to these APCD records</td>
<td></td>
</tr>
<tr>
<td>4455544</td>
<td>ABCD</td>
<td>HUJKLMNOPQ</td>
<td>POIUYT</td>
<td>02116</td>
<td>F</td>
<td>4 Matches, 1 Mismatch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4455544</td>
<td>ABCD</td>
<td>QRSTUVWXYZ</td>
<td>POIUYT</td>
<td>02116</td>
<td>M</td>
<td>4 Matches, 1 Mismatch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4455544</td>
<td>MNOP</td>
<td>QRSTUVWXYZ</td>
<td>POIUYT</td>
<td>02116</td>
<td>F</td>
<td>4 Matches, 1 Mismatch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2332332</td>
<td>ABCD</td>
<td>QRSTUVWXYZ</td>
<td>LKJHGD</td>
<td>02116</td>
<td>F</td>
<td>4 Matches, 1 Mismatch, DOB weighted stronger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4455544</td>
<td>ABCD</td>
<td>HUJKLMNOPQ</td>
<td>POIUYT</td>
<td>02116</td>
<td>M</td>
<td>3 Matches, 1 Mismatch</td>
<td>Lower</td>
<td>Based on Study Requirements, Input Row may link to these APCD Records</td>
<td></td>
</tr>
<tr>
<td>5755542</td>
<td>ABCD</td>
<td>MNBCDVSWX</td>
<td>LKJHGD</td>
<td>02119</td>
<td>F</td>
<td>2 Matches, 3 Mismatch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7886655</td>
<td>MNOP</td>
<td>HUJKLMNOPQ</td>
<td>POIUYT</td>
<td>02116</td>
<td>M</td>
<td>2 Matches, 3 Mismatch</td>
<td>Too Low</td>
<td>Input Row does not link to these APCD records</td>
<td></td>
</tr>
</tbody>
</table>
Example Matching Projects

Successful data linkage projects leveraging *pseudonymized* identifiers

- Dept. of Public Health study linking to opioid data (CH. 55)
- Dept. of Public Health *Public Health Data Warehouse* (included linking of 21 datasets)
- Dept. of Elder Affairs study linking long-term services and support data & federal Housing & Urban Development housing data
- Dept. of Public Health study linking to birth certificate records to study postpartum depression
- Dept. of Public Health studying linking to assisted reproductive technology data

In Progress

- Dept. of Public Health study linking public housing and smoking cessation data
- U.S. Dept. of VA study linking to VA hospital data
- Brigham and Women’s study linking to cardiac data
For questions, please contact:

- Kathy Hines
  - Kathy.Hines@state.ma.us (617) 701-8275
- Scott Curley
  - Scott.Curley@state.ma.us (617) 701-8255