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

OVERVIEW

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

Analytic Challenge

- 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 (MMYYYY 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

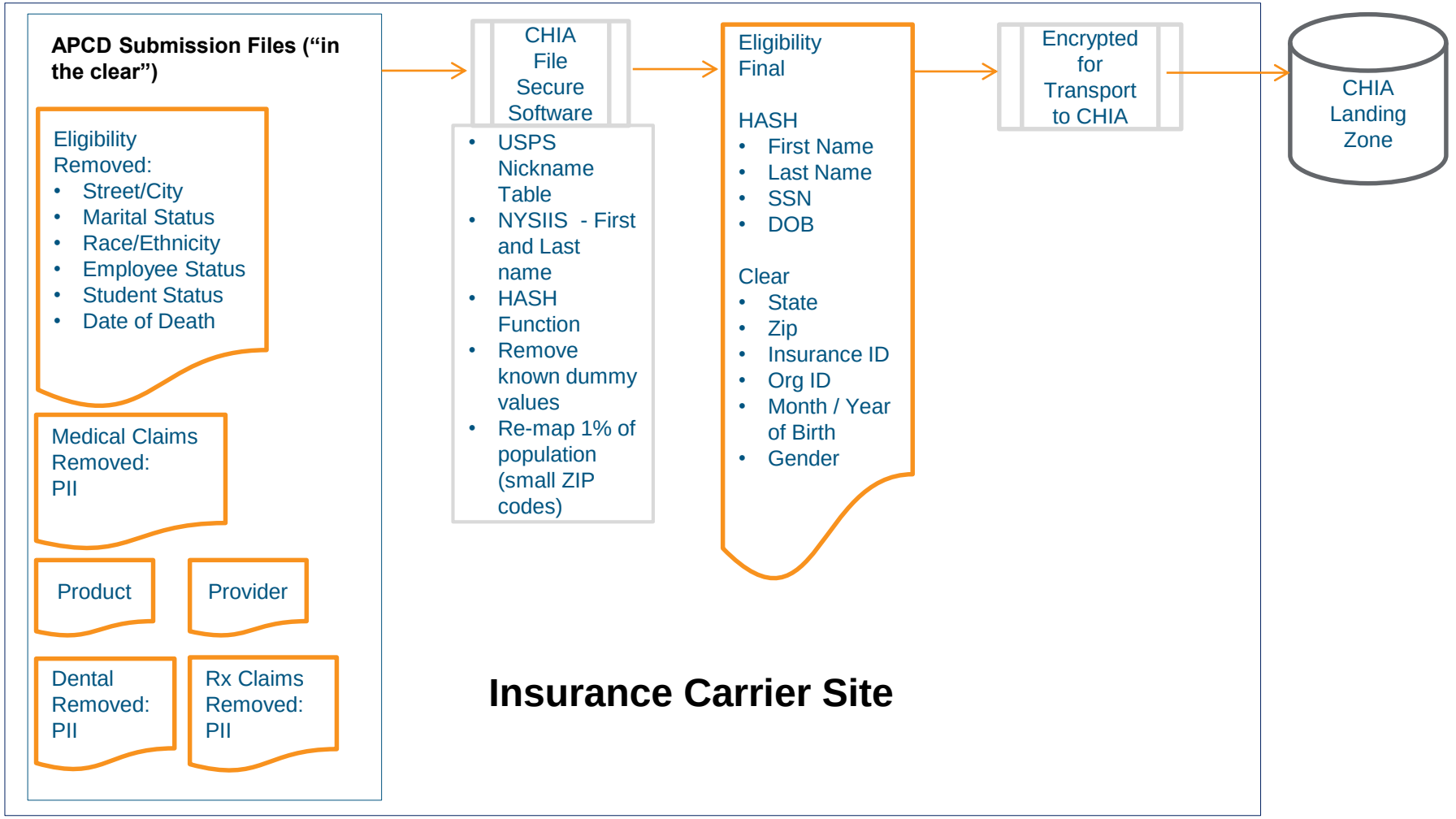
- Zip code processing
  - Flag if invalid zip code
  - Retain MA Zip codes only
  - Map MA Zip codes to mask small areas in MA APCD
- State code processing
  - Flag if invalid state
  - Retain only New England and New York state codes
  - Map MA Zip codes to mask small areas in MA APCD
- *File Secure*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

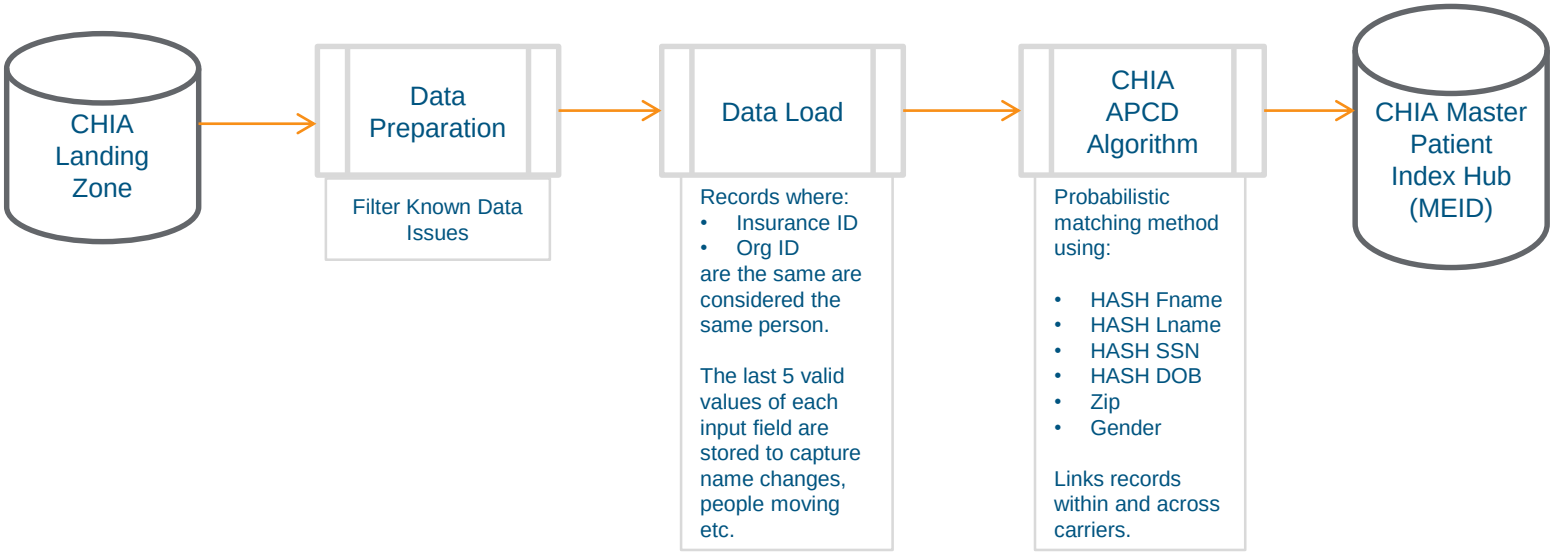


MASTER PATIENT INDEX (MPI)

MPI and Record Linking

- 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

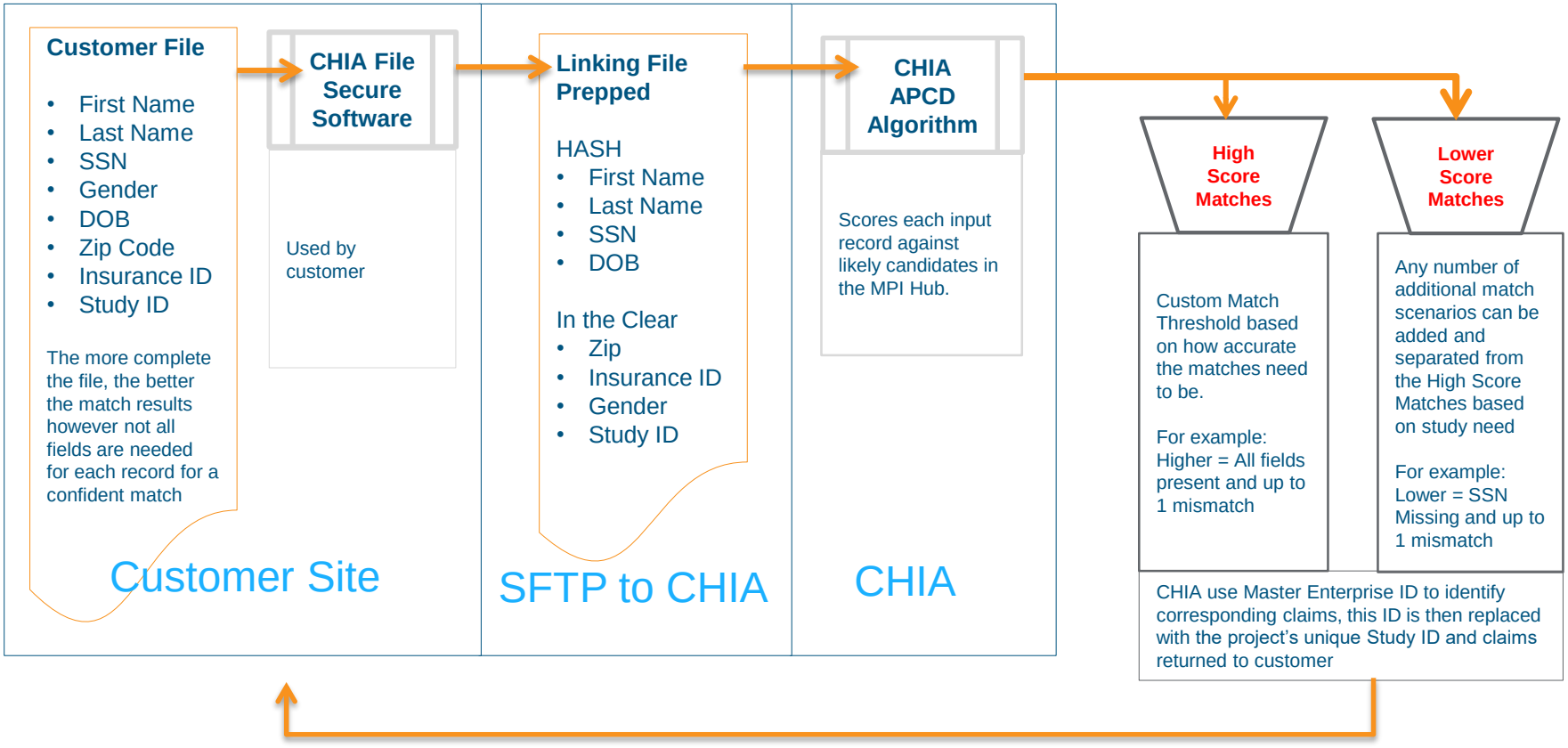
CHIA Master Patient Index



CHIA MPI	Org ID	Insurance ID	First Name	Last Name	DOB	Gender	SSN	ZIP Code
111111	30	BBY00002211	ABCD	QRSTUVWXYZ	POIUYT	F	HFHDSFH	02116
				KDFGJKDFKFK				02461
								02090
112233	22	HVD00000122	QWDD	DGFGDFFGFG	GFGDFF	M	FGDDDFG	02118
112233	30	BBY000034234	QWDD	DGFGDFFGFG	GFGDFF	M	FGDDDFG	01056
								13116
								01025

CHIA MATCHING SERVICE

CHIA Matching Service (Master Data Management)



CHIA Linkage Service (MPI Search)

Input Row from Customer - Hashed Equivalent

Study ID	First Name	Last Name	DOB	SSN	Zip Code	Gender
8888	ABCD	QRSTUVWXYZ	POIUYT		02116	F

APCD Linking Scenarios

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

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 cessastion data
- U.S. Dept. of VA study linking to VA hospital data
- Brigham and Women's study linking to cardiac data

Contact Information

For questions, please contact:

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