DATAVANT

Matching Accuracy of Patient Tokens in De-Identified Health Data Sets

A False Positive Analysis

Executive Summary

One of the most important and early tasks all healthcare analytics organizations face is the need to protect private personal information. This task is made harder by the need to establish an adequate understanding of an individual's or a group's health care status by combining disparate data from multiple sources. Encrypted patient tokens allow matching of patient records across separate data sets without exposure of the underlying protected health information (PHI). This study assessed the matching accuracy of two common token types to understand how many matches were unique, and how many were false positives.

Key findings include:

- Tokens built from the combination of the first initial of the first name, last name, date of birth, and gender allow 96.3% accurate matching, and generate 3.7% false positive matches
- Tokens built from the combination of the Soundex of first and last name, date of birth, and gender allow 96.1% accurate matching, and generate 3.9% false positive matches
- Using both tokens together allows 98.9% accurate matching, with only 1.1% false positive matches

De-identification of health data: Protecting privacy to enable Big Data analytics in healthcare

Big data analytics in healthcare has long been a goal for providers, payers, and biopharma manufacturers, but important barriers have impeded progress. The most common barriers in the United States are regulatory, predominantly outlined in restrictions set forth in the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and in the subsequent Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009. These laws outlined the necessary provisions to encourage use of health information, but they also stipulated the security and privacy protections that need to be followed by anyone hoping to use healthcare data for big data analysis.

HIPAA in particular stipulates the protected health information (PHI) elements that need to be removed from a healthcare data set to be considered de-identified. In short, de-identified health data can be created using the HIPAA Safe Harbor method, whereby one removes all information falling into 18 different categories (e.g. names, addresses, dates except years, phone numbers, etc.). Alternatively, health data users can use the statistical method to remove less information, but always enough to make it statistically impossible to re-identify the underlying patient. Statistical de-identification methods always remove names, addresses, and other personally identifiable information (PII), but are often able to leave

important analytical elements, including dates of service and 3-digit zip codes in de-identified health data sets.

Regardless of the method used, the outcome is that the individual patient record is de-identified – or "anonymized". Unfortunately, this anonymization means that two de-identified health data sets cannot be merged together because it is impossible to identify and match one patient's record in one data set with their records in another data set. Datavant has solved this problem through development of software that, as it performs HIPAA-compliant de-identification on the underlying data set, also inserts a unique encrypted patient token into each record. These patient tokens are reliably and reproducibly created in any health data set, such that the same token is created for the same patient wherever the software is run. In this way, users can join de-identified health data sets for big data health analytics by matching the encrypted patient tokens from one record to another. But how accurate is this matching process?

Using encrypted patient tokens to merge de-identified health data: a study in matching accuracy

To understand the matching accuracy of the two encrypted patient token designs most commonly used by Datavant clients, we performed an analysis of how often a patient token scheme uniquely matched a patient in a population-wide data set.

Test data set

To perform this analysis, we used the data file underlying our mortality data module, which provides mortality data for the United States based on the information reported in the Death Master File (DMF) from the Social Security Administration, complemented by data gathered from obituaries since 2010. This data file contains the names, genders, and birthdays for almost 100 million people across the United States. Beyond the very large sample size, we chose this file as our test data set because it is not biased to any geography or other demographic.

We assume that there are individuals in the United States with the same name, gender, and birthdate (indeed, this analysis was built to quantify this overlap or "non-uniqueness"), and the breadth of this data set is large enough that these non-unique individuals should be present in large numbers. Importantly, we can use the presence of distinct social security numbers (from the DMF) to prove that people with this same PII are actually distinct individuals. Likewise, we can use the presence of different dates of death in the obituary data to prove that people with this same PII in that data set are also actually distinct individuals.

Filling in missing data

Like many big health data sets, the original DMF and obituary data files are incomplete in reporting all of the fields we would like for this analysis – particularly gender. Therefore, we first added a gender to missing records. To determine the likely gender of an individual in the data set, we compared the first name in each record against a large consumer list that reported both first names and gender. Looking at the percentage of individuals for whom a name matched a particular gender, we determined the likelihood that each individual in our test file was a certain gender (e.g. David is almost exclusively associated with a male gender, whereas Sam or Chris are more mixed because they are abbreviations for Samuel or Samantha, or Christopher or Christina, respectively). Every person in our test data file for whom we had a gender likelihood greater than 50% (i.e. at least 50% of the people with that first name were that gender) was included in the final test data set for this analysis.

Test patient tokens:

Encrypted patient tokens created by the Datavant de-identification engine are generated from the underlying PII in the health data set (before it is de-identified). For this study, we used two Datavant token schemes that incorporate the following fields:

- Patient Token 1: Last Name + First_Initial + Gender + date of birth (DOB)
- Patient Token 2: Last Name (Soundex) + First Name (Soundex) + Gender + DOB

These two token schemes are the most commonly used across our Datavant client base because most healthcare (and other) data sets have these fields. Additionally, these token schemes allow some degree of "fuzzy" matching in that Token 1 only uses the first initial, allowing names that are commonly abbreviated to be matched (e.g. Chris, Christy, and Christina), and Token 2 uses the Soundex principle which corrects for misspellings in names. (Note that these tokens support probabilistic matching, and we recommend using deterministic tokens – those based on unique identifiers like social security numbers – wherever possible.)

Test for matching accuracy using two common encrypted patient tokens

The Datavant de-identification engine was used to create both Token 1 and Token 2 for every individual in the final test data set, creating a record set of >380 million tokens. (Note that because gender is not a field in the original data set, we generated a token for both genders in a number of cases, such that the number of tokens exceeds the number of individuals.)

If a token is only found once in the entire record set, it can be reasonably concluded that it represents a unique combination of the PII fields that went into its creation (i.e. no one else has that combination of name, date of birth, and gender). Alternatively, if a token is found multiple times in the record set, then

it can be reasonably concluded that multiple individuals share the PII fields that created it.

If a patient token is determined to be unique, one can also conclude that any match using this particular patient token in de-identified health is an accurate match. However, a match across de-identified health data sets using a token that is shared by multiple individuals would be considered a potential false positive, in that one could not be sure that the two matching records actually belonged to the same individual. In fact, one should assume that multiple patients are represented in the matching record set of a non-unique patient token.

Therefore, it is critical to understand the uniqueness of each patient token in order to understand the matching accuracy of the patient token scheme. To perform this analysis, we counted the number of times each patient token appeared in our >380 million token set, and reported the results below.

Patient token uniqueness (and expected match accuracy rates):

Token 1 Uniqueness

As stated above, Token 1 is created using the first initial, full last name, date of birth, and gender. Therefore, for example, if John Smith and Justin Smith have the same birthday, they will share the same Token 1.

In our data set, there were 142.3 million different Token 1s. Of these, the vast majority (96.3%) mapped to just a single record, meaning they were unique for that individual. 3.1% of Token 1s were shared by two different records (i.e. shared by 2 different people). As expected, even fewer Token 1s were shared by 3 individuals, and fewer still were shared by more than that. See Table 1 for the full results.

Table 1: Record match rates (uniqueness) when using Token 1

Number of records with each Token 1	Count of Token 1s	Rate of Uniqueness
1 (completely unique)	137,072,611	96.31%
2 records share token	4,374,571	3.07%
3 records share token	627,949	0.44%
4 records share token	164,765	0.12%
5 records share token	54,223	0.04%
6 records share token	19,713	0.01%
7 records share token	7,249	0.01%
8 records share token	2,787	0.00%
9 records share token	1,027	0.00%
10+ records share token	583	0.00%
Total	142,325,478	100%

Token 2 Uniqueness

Token 2 is created using the Soundex of the full first name and last name, date of birth, and gender. Therefore, remembering that the Soundex algorithm standardizes homophones, if John Smith and Jon Smythe have the same birthday for example, they will share the same Token 2.

In our data set, there were 142.8 million different Token 2s. (Note that there are slightly more unique Token 2s created than Token 1s because using only a first initial in Token 1 is not quite as discriminatory of different names.) Of these different Token 2s, 96.1% mapped to just a single record, which is similar to what we saw with Token 1. See Table 2 for the full results. Though the differences are small, we can see that Token 2 creates slightly more unique matches than Token 1.

Table 2: Record match rates (uniqueness) when using Token 2

Number of records with each Token 2	Count of Token 2s	Rate of Uniqueness
1 (completely unique)	137,240,429	96.11%
2 records share token	5,116,869	3.58%
3 records share token	380,425	0.27%
4 records share token	44,788	0.03%
5 records share token	7,260	0.01%
6 records share token	1,353	0.00%
7 records share token	339	0.00%
8 records share token	80	0.00%
9 records share token	23	0.00%
10+ records share token	25	0.00%
Total	142,791,591	100%

Combining Token 1 and Token 2 for greater matching accuracy

As both Token 1 and Token 2 allow fuzzy matching as described in the Test Patient Tokens section above, it is unsurprising that they do not generate perfect uniqueness rates of unique tokens in this analysis. However, because they approach fuzzy matching in fundamentally different ways, we assessed whether the combination of the two tokens would identify a unique individual with even greater accuracy than when used alone.

As shown in Table 3 below, the combination of Token 1 and Token 2 showed a substantial increase in uniqueness in the record set. The combination of Token 1 and Token 2 define a unique individual (only one instance of the combination in the entire record set) nearly 99% of the time. 1% of the time, there are two individuals who share the same combination of Token 1 and Token 2. According our analysis, only 0.07% of individuals could be confused with 2 or more other individuals when using the combination

of Token 1 and Token 2.

Table 3: Record match rates when using the combination of Token 1 and Token 2

Number of records with each Token 1+2 Combination	Count of Token 1+2 combinations	Rate of Uniqueness
1 (completely unique)	145,522,099	98.91%
2 records share token	1,508,403	1.03%
3 records share token	83,186	0.06%
4 records share token	10,549	0.01%
5 records share token	1,790	0.00%
6 records share token	330	0.00%
7 records share token	89	0.00%
8 records share token	21	0.00%
9 records share token	10	0.00%
10+ records share token	2	0.00%
Total	147,126,479	100%

Study conclusions: combining probabilistic patient tokens to allow high accuracy matching of de-identified health data

Token 1 and Token 2 are a powerful combination for generating unique matches of individuals across data sets. There is a false positive rate of slightly less than 1% when using these tokens together, meaning that a match of patient records using the combination of Token 1 and Token 2 may not indicate that the correct patient records are linked even though the tokens match.

To reduce the false positive rate even more, we recommend using other fields like zip code or national provider identifier (NPI) numbers for providers as additional verification that a match is indeed for the same individual. Likewise, users can also generate additional tokens including the full first name and other variations of the underlying PII to increase the accuracy of the matching process. Where possible, we always recommend using deterministic tokens (those based on truly unique PII like social security numbers) for matching where the data sets have the information to support it.

For more information:

- Contact Jason LaBonte, Ph.D. for questions or comments about this analysis:
  Jason@datavant.com

- Contact Lauren Stahl for more information about the Datavant modules that were used in this study:
  Lauren@datavant.com

- Visit the Datavant website to read our other whitepapers and materials:
  www.datavant.com

Organizing the World's Health Data

Datavant helps organizations safely share and link healthcare data.

We believe in connecting healthcare data to eliminate the silos of healthcare information that hold back innovative medical research and improved patient care. We help data owners manage the privacy, security, compliance, and trust required to enable safe data sharing.

Datavant's vision is backed by Roivant Sciences, Softbank, and Founders Fund, and combines technical leadership and healthcare expertise. Datavant is located in the heart of San Francisco's Financial District.

Glossary of Terms:

Covered Entity

A covered entity (CE) under HIPAA is a health care provider (e.g. doctors, dentists, pharmacies, etc), a health plan (e.g. private insurance, government programs like Medicare, etc), or a health care clearinghouse (i.e. entities that process and transmit healthcare information).

De-identified health data

De-identified health data is data that has had PII removed. Per the HIPAA Privacy Rule, healthcare data not in use for clinical support must have all information that can identify a patient removed before use. This rule offers two paths to compliantly remove this information: the Safe Harbor method and the Statistical method. When these identifying elements have been removed, the resulting de-identified health data set can be used without restriction or disclosure.

Deterministic matching

Deterministic matching is when fields in two data sets are matched using a unique value. In practice, this value can be a social security number, Medicare Beneficiary ID, or any other value that is known to only correspond to a single entity. Deterministic matching has higher accuracy rates than probabilistic matching, but is not perfect due to data entry errors (mis-typing a social security number such that matching on that field actually matches two different individuals).

Encrypted patient token

Encrypted patient tokens are non-reversible 44 character strings created from a patient's PHI, allowing a patient's records to be matched across different de-identified health data sets without exposure of the original PHI.

False positive

A false positive is a result that incorrectly states that a test condition is positive. In the case of matching patient records between data sets, a false positive is the condition where a "match" of two records does not actually represent records for the same patient. False positives are more common in probabilistic matching than in deterministic matching.

Fuzzy matching

Fuzzy matching is the process of finding values that match approximately rather than exactly. In the case of matching PHI, fuzzy matching can include matching on different variants of a name (Jamie, Jim, and Jimmy all being allowed as a match for "James"). To facilitate fuzzy matching, algorithms like SOUNDEX can allow for differently spelled character strings to generate the same output value.

Health Information Technology for Economic and Clinical Health (HITECH) Act

The HITECH Act was passed as part of the as part of the American Recovery and Reinvestment Act of 2009 (ARRA) economic stimulus bill. HITECH was designed to accelerate the adoption of electronic medical records (EMR) through the use of financial incentives for "meaningful use" of EMRs until 2015,

and financial penalties for failure to do so thereafter. HITECH added important security regulations and data breach liability rules that built on the rules laid out in HIPAA.

Health Insurance Portability and Accountability Act of 1996 (HIPAA)

HIPAA is a U.S. law requiring the U.S. Department of Health and Human Services (HHS) to develop security and privacy regulations for protected health information. Prior to HIPAA, no such standards existed in the industry. HHS created the HIPAA Privacy Rule and HIPAA Security Rule to fulfill their obligation, and the Office for Civil Rights (OCR) within HHS has the responsibility of enforcing these rules.

Personally-identifiable information (PII)

Personally-identifiable information (PII) is a general term in information and security laws describing any information that allows an individual to be identified either directly or indirectly. PII is a U.S.-centric abbreviation, but is generally equivalent to "personal information" and similar terms outside the United States. PII can consist as informational elements like name, address, social security number or other identifying number or code, telephone number, email address, etc., but can include non-specific data elements such as gender, race, birth date, geographic indicator, etc. that together can still allow indirect identification of an individual.

Probabilistic matching

Probabilistic matching is when fields in two data sets are matched using values that are known not to be unique, but the combination of values gives a high probability that the correct entity is matched. In practice, names, birth dates, and other identifying but non-unique values can be used (often in combination) to facilitate probabilistic matching.

Protected health information (PHI)

Protected health information (PHI) refers to information that includes health status, health care (physician visits, prescriptions, procedures, etc.), or payment for that care and can be linked to an individual. Under U.S. law, PHI is information that is specifically created or collected by a covered entity.

Safe Harbor de-identification

HIPAA guidelines requiring the removal of identifying information offered covered entities a simple, compliant path to satisfying the HIPAA Privacy Rule through the Safe Harbor method. The Safe Harbor de-identification method is to remove any data element that falls within 18 different categories of information, including:

1. Names
2. All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes. However, you do not have to remove the first three digits of the ZIP code if there are more than 20,000 people living in that ZIP code.
3. The day and month of dates that are directly related to an individual, including birth date, date of admission and discharge, and date of death. If the patient is over age 89, you must also remove his age and the year of his birth date.

4. Telephone number
5. Fax number
6. Email addresses
7. Social Security number
8. Medical record number
9. Health plan beneficiary number
10. Account number
11. Certificate or license number
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Web addresses (URLs)
15. Internet Protocol (IP) addresses
16. Biometric identifiers, such as fingerprints
17. Full-face photographs or comparable images
18. Any other unique identifying number, such as a clinical trial number

Social Security Death Master File

The U.S. Social Security Administration maintains a file of over 86 million records of deaths collected from social security payments, but it is not a complete compilation of deaths in the United States. In recent years, multiple states have opted out of contributing their information to the Death Master File and its level of completeness has declined substantially. This Death Master File has limited access, and users must be certified to receive it. This file contains PHI elements like social security numbers, names, and dates of birth – therefore, bringing the raw data into a healthcare data environment could risk a HIPAA violation.

Soundex

Soundex is a phonetic algorithm that codes similarly sounding names (in English) as a consistent value. Soundex is commonly used when matching surnames across data sets as variations in spelling are common in data entry. Each soundex code generated from an input text string has 4 characters – the first letter of the name, and then 3 digits generated from the remaining characters, with similar-sounding phonetic elements coded the same (e.g. D and T are both coded as a 3, M and N are both coded as a 5).

Statistical de-identification (also known as Expert Determination)

Because the HIPAA Safe Harbor de-identification method removes all identifying elements, the resulting de-identified health data set is often stripped of substantial analytical value. Therefore, statistical de-identification is used instead (HIPAA calls this pathway to compliance "Expert Determination"). In this method, a statistician or HIPAA certification professional certifies that enough identifying data elements have been removed from the health data set that there is a "very small risk" that a recipient could identify an individual. Statistical de-identification often allows dates of service to remain in de-identified data sets, which are critical for the analysis of a patient's journey, for determining an episode of care, and other common healthcare investigations.