Overview of Datavant's De-Identification and Linking Technology for Structured Data

Introduction

Datavant is firmly committed to advancing healthcare through data analytics while protecting patients' privacy. Before Datavant, private patient information was protected by removing it altogether. If it wasn't there, the thinking went, it couldn't be exposed. Unfortunately, if a patient's identifying information was removed, it also meant that there was no way to combine that patient's healthcare data in one file (e.g., his or her hospital stay) with data in another file (e.g., his or her pharmacy prescriptions after being discharged).

At Datavant, we take a different approach. We've designed cutting-edge, patent-pending, de- identification technology that replaces private patient information with an encrypted "token" that can't be reverse-engineered to reveal the original information. Furthermore, our technology can create these same patient-specific tokens in any data set, which means that now two different data sets can be combined using the patient tokens to match corresponding records without ever sharing the underlying patient information.

Datavant's technology is installed and run locally behind the users' firewalls. No patient information is sent out to Datavant, and Datavant does not have access to users' systems. Installing and running Datavant's software is simple, fast and easy, and the technology is currently installed at providers, payers, life sciences companies, analytics companies, and large data aggregators.

Rapid and Repeatable De-Identification

Datavant's de-identification engine is designed specifically for use on structured healthcare data (though it can operate on any structured data set). The de-identification engine performs two functions: (i) de-identification of the data set (including removal of patient information and modifications of patient information), and (ii) the insertion of encrypted patient tokens.

Datavant's technology creates statistically de-identified data sets by removing personally-identifiable information (PII; which includes but is not necessarily limited to protected health information, or PHI) like names and medical record numbers; and modifying other values, such as turning 5-digit zip codes into 3-digit zip areas, or converting dates of birth to years of birth.

As the technology de-identifies a patient record, it also generates one or more tokens for that record. These tokens are based on the PII in the record. As a result, the tokens can be consistently created from any data set where the underlying PII is the same. Matching tokens can be used to link a patient's record in one data set with a record for the same patient in a different set, *without ever exposing the PII of that patient.*

Tokens can be built from many different combinations of PII elements, and the specific tokens to be created will be specified during the configuration process. Multiple tokens are often created to facilitate matching. Tokens can be built based on fields such as social security number, which allows for deterministic matching; alternately, they can be constructed from fields such as name and date of birth, which in combination can be used to support probabilistic matching. Over years of implementations, Datavant's QA testing protocols have shown that Datavant's technology generates reliable, repetitive tokens.

Configuration of the De-Identification Process

Datavant's technology can be configured to work with *any* data layout, and to incorporate the specific identification rules that the data owner would like to use. The configuration process is a collaborative effort between Datavant and the data owner to ensure that the result satisfies the data owner's regulatory and business compliance needs.

Configuration starts with Datavant working with the data owner to define the format of the input data to the de-identification engine. Input data can be defined by either the use of standard formatting instructions (e.g., 837 medical claims, NCPDP pharmacy claims, HL7 ADT messages, etc.) or by joint design efforts with the data owner's technical team. For example, the team may specify pipe delimited database extracts that will serve as the inputs. Once the format has been defined, Datavant and the data owner identify all PII data elements and their positions. This includes identification of non-obvious PII elements, such as the location of a service address that might be identified as "home service". These PII elements are what will be removed or modified by Datavant's technology, according to the rules defined in the software's "template" file.

Once complete, the template contains all of the de-identification rules to be applied when the technology is run by the data owner. Datavant can configure the software to de-identify data sets to completely remove PHI in

order to allow data owners to comply with the Safe Harbor method outlined in the Health Insurance Portability and Accountability Act (HIPAA). More frequently, however, data owners seeking to preserve the analytical value of their data sets choose rules that allow statistical de-identification per the Expert Determination method outlined in HIPAA. To support the Expert Determination methodology, Datavant frequently removes PHI values like name, date, gender, and patient and zip code. Table 1 describes the rules typically applied to each HIPAA-designated PHI element under the Expert Determination methodology.

Note that Datavant does not shift the patient's dates of service, as expert HIPAA certifiers do not believe date-shifting is an effective de-identification methodology. Furthermore, date-shifting makes it very difficult to perform analytics on longitudinal patient records linked across different organizations.

Table 1: Common De-Identification Rules Datavant Applies to Protected Health Information (PHI)

Names	Removed where present
Zip Code	All patient and subscriber zip codes are reduced to the initial three characters to define a zip area. Based on HIPAA rules, however, even three-digit zip areas with a combined population of less than 20,000 are either nulled out or are combined again with additional zip zones to ensure that populations exceed this minimum
Date of Service	Dates of service (e.g., admission dates, discharge dates, prescription fill dates, and procedure dates) are typically preserved when using a statistical de- identification methodology
Date of Birth	The standard template's birthday rule converts all birth dates to birth year (the data source compliance officer may alter this to month instead of year). All dates of birth where the individual would be 89 years of age or greater as of the date of de-identification would be modified to reflect an age of 89
Medical Records Numbers	Removed where present
Telephone Numbers	Removed where present
Email Address	Removed where present
Social Security Numbers	Removed where present
Beneficiary Numbers	Removed where present
Vehicle Information	Removed where present
Device Identifiers and Serial Numbers	Removed where present
URL Addresses	Removed where present
IP Addresses	Removed where present

Biometric Values	Removed where present
Image Fields	Removed as defined by the data source

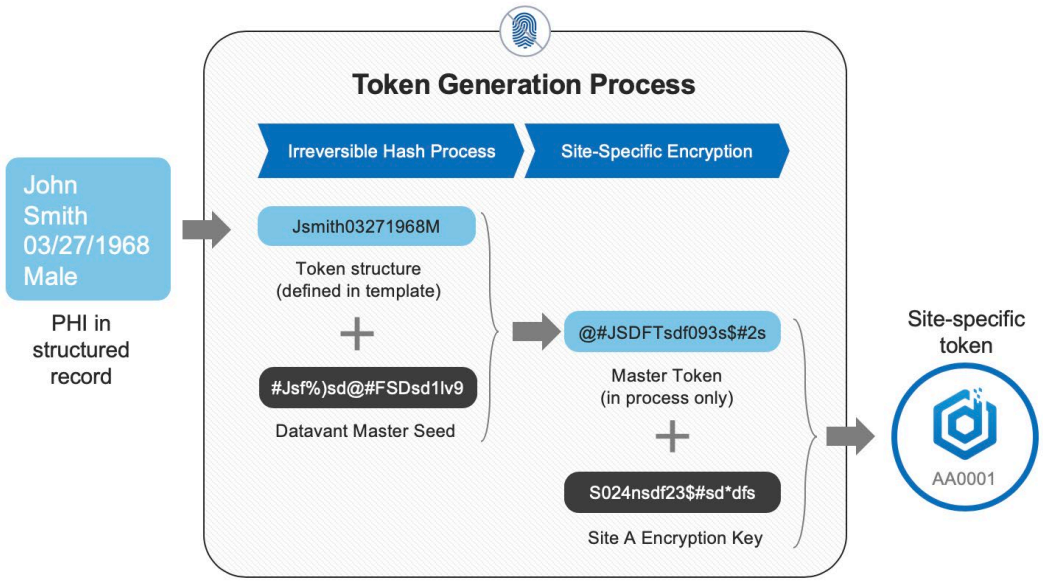
Creation of Site-Specific, Encrypted Patient Tokens

The same template that is used to specify the data owner's de-identification rules is also used to specify the PII values that will be used to generate tokens. In the process of tokenization, those PII values will be hashed and encrypted. Tokens are used to identify and link matching individual records across different data sets without ever exposing the PII of the patient to whom each record belongs.

Datavant has developed a unique method to ensure that encrypted tokens are (i) irreversible, and (ii) site-specific, meaning that each Datavant user's tokens are unique for that user (patent pending).

In order to be certified under HIPAA, it is critical that the tokens used to link records in de-identified data sets cannot be reversed to reveal the patient's identifying information. The first step of the token creation process (see Figure 1) is the use of an **irreversible hash function**, which ensures that the patient's PII used to create the token cannot be recovered from the output value. In the second step of the process, the hash value (i.e., "Master Token") is encrypted with a **site-specific encryption** key to generate the final encrypted patient token. While the same patient information will always create the same Master Token, site-specific encryption means that a single patient will have a unique token (i.e., a "site-specific token") in each specific user's data set. Site-specific encryption ensures that a breach at one data site will never compromise the tokens or PHI at any other data site.

Figure 1: Datavant creates irreversible, site-specific encrypted tokens for each patient record



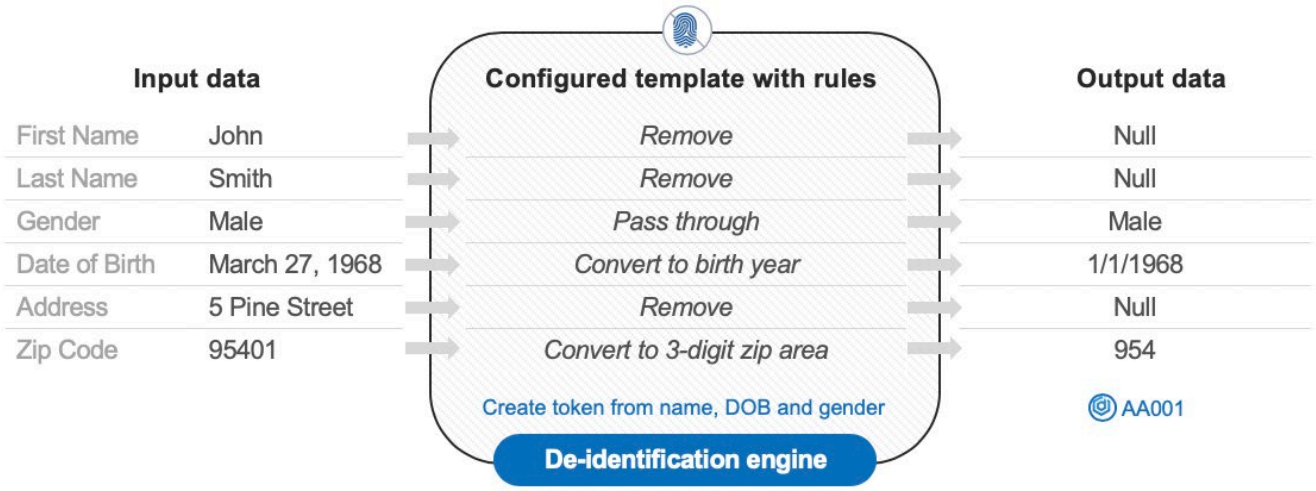
The PII elements used in the token creation process are defined in the template file. Datavant works with data owners to determine which fields should be used to create tokens, a choice which is driven by which PII fields exist in the data sets that a data owner would like to link to.

Datavant recommends that every template specify name, date of birth, and gender as elements to be used in the token generation process. These fields are present in almost every data set and the tokens that can be constructed from these elements generate high rates of matching accuracy. Additional tokens can be generated for each record using other PII elements to facilitate matching across a variety of data sets (some of which the data user may not yet know that they would like to link to).

Using the Configured Template to Generate De-Identified Data Files

Once the de-identification and token creation rules are defined and mapped to the appropriate PII elements in the input data set, the final template is used to configure Datavant's software (see Figure 2). Note that the Master Token is *never* present in any output or log stream; only the site-specific encrypted tokens are written to the output file.

Figure 2: Use template to configure Datavant's software, which reliably generates a de-identified and tokenized output file



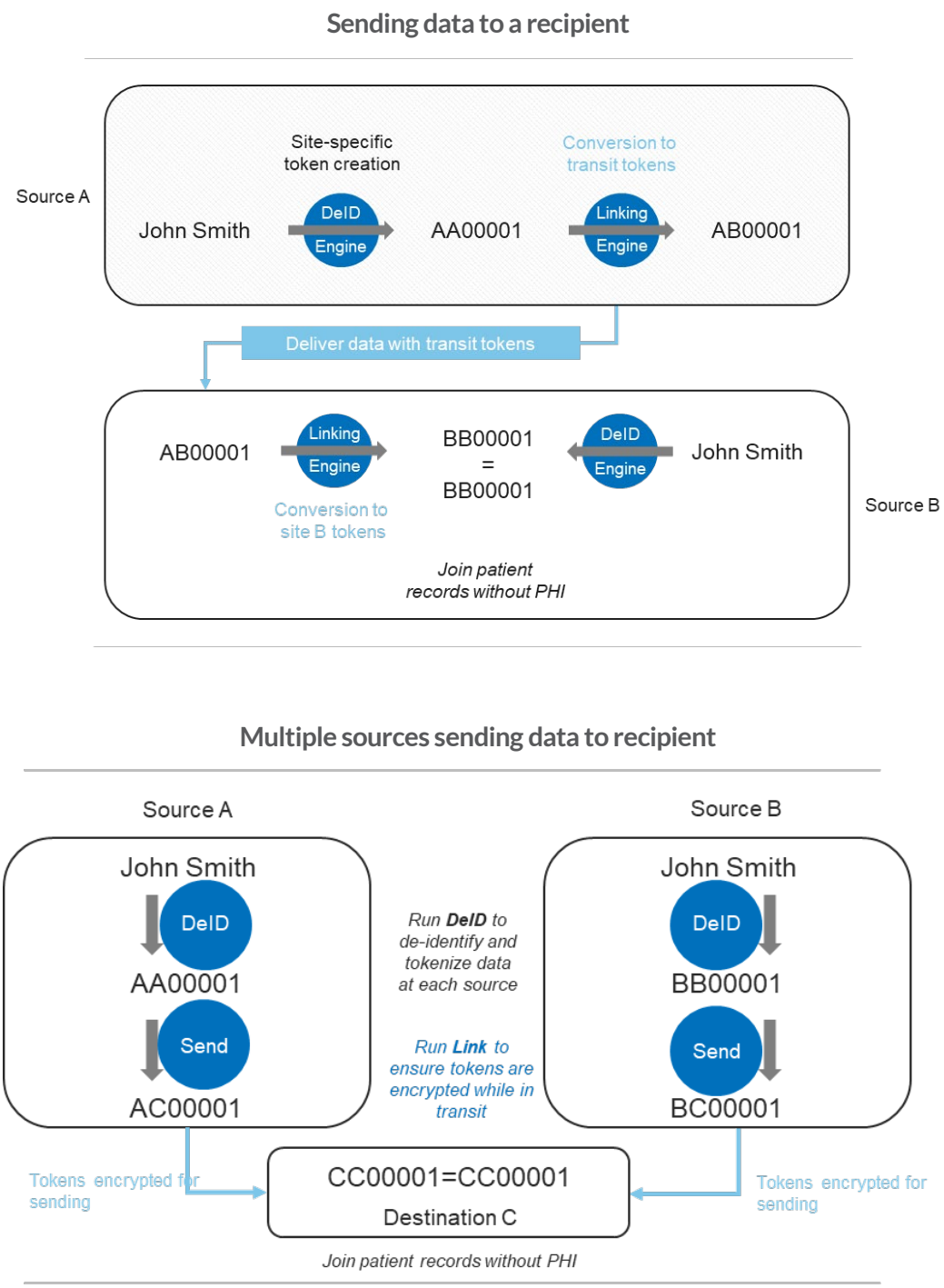
Token Transformation to Allow Matching Patient Records

As discussed above, Datavant's software creates de-identified data sets with site-specific tokens for each patient record. For many data users, the next step is to merge disparate data sets. In order to do so, the site-specific tokens from each party are *transformed* into the same token scheme, allowing matching patient records to be identified.

Like our structured data de-identification software, our token transformation (or linking) software is installed and run locally at the site. It allows a data source to convert the site-specific tokens created during the de-identification process into a data recipient's site-specific token scheme. As a result, corresponding de-identified patient records in the data source and recipient's data can now be matched.

In practice, there is an intermediate step to converting a data source's token scheme to a data recipient's token scheme. The source first transforms their site-specific tokens to *transit tokens*, which ensures that the source's site-specific tokens never leave the source's environment. The data is sent with encrypted transit tokens, and then is transformed again by the recipient into their token scheme. See Figure 3 below for an illustration of the token transformation process that enables matching of patient records.

Figure 3: Datavant tokens allow corresponding patient records to be matched across data sets without ever sharing PHI



The primary feature of this token transformation process is that *the patient key is never exposed*. What this means is that if a site token value were ever to be exposed to another party, (i) the exposure only impacts that particular site, and (ii) Datavant can simply provide the site a new site key and new software. Existing data sets can be de-identified again using the new key and software, eliminating the exposure created by the loss of the original site token values.

In addition to increasing the level of protection against PHI violations, the site-level token system allows the data source to control the usage of its data by other organizations. To transform tokens from one site key to another site key, Datavant's technology requires permission from that other site. Thus, data sources are able to exercise ongoing control over the link-ability of their data.

Flexible Usage of the Technology

Datavant's technology is delivered to the environment where data is being processed, whether at individual data sources or at centralized data aggregators. The software is installed and run at the site itself – *no PII is required to be sent out from the site* – and Datavant has no access to the site's systems.

As discussed above, the software can be easily configured to support any data layout based on the configuration rules defined in the template, which also allows for custom modification and removal rules based on a data user's de-identification needs. Further, users can create multiple templates in their first implementation, and then select the rules they would like to be applied to specific data sets going forward.

How Our Clients Take Advantage of Datavant's Technology

Datavant is installed at sites across the healthcare spectrum. All healthcare stakeholders face the common challenge of protecting patient privacy while maximizing their data's utility for healthcare analytics. Datavant offers a simple, reliable, and flexible way to de-identify data sets in a way that can be deemed adequate under HIPAA while still retaining the ability to link data sets from multiple sources without exposure of PHI. Here is how different stakeholders are using Datavant today:

Healthcare Providers: Clinicians and facilities providing direct patient care.

- Datavant offers healthcare providers that manage large volumes of patient data the ability to reduce risk under HIPAA. Datavant protects against regulatory violations by employees or

Business Associates who do not have a legitimate need to access PHI.

Healthcare Payers: Organizations that pay for healthcare services.

- Datavant offers healthcare payers the simplest way to link and analyze healthcare data in a way that can be deemed adequate under HIPAA.

Healthcare Service Providers: HIPAA-covered entities or their Business Associates (entities that have signed a Business Associate Agreement, or BAA) who support healthcare delivery (patient care, payment, and related services).

- Datavant supports the efforts of healthcare service providers to link, exchange, or sell healthcare data to multiple organizations safely and securely, while reducing HIPAA-compliance risk.

Pharmaceutical and Medical Device Manufacturers, Data Aggregators and Analytics Companies: Each of these types of organizations analyze healthcare transaction data but are not equipped to manage the responsibility required of a HIPAA entity or a Business Associate.

- Datavant offers the broader healthcare ecosystem the ability to discover greater insights from healthcare transaction data aggregated from multiple sources in a HIPAA-compliant manner.

Non-Profit or Academic Researchers: Organizations that conduct research using patient or healthcare transaction data to address questions of policy, to advance understanding of health service delivery or patient outcomes.

- Datavant offers a simple and affordable means of aggregating de-identified healthcare transaction data from disparate sources in a manner that can be certified by an expert under HIPAA's expert determination method, addresses institutional review board (IRB) requirements, and meets rigorous scientific standards.

For more information:

- Contact Bob Borek, Head of Marketing (bob@datavant.com) for questions or comments about this analysis.
- Visit the Datavant website to read our other white papers and materials (www.datavant.com).

Connecting the World's Health Data

Datavant helps organizations safely protect, match and share health 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 exchange.

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, or pharmacies), a health plan (e.g., private insurance, or government programs like Medicare), 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 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 (e.g., 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 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 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 of informational elements like name; address; social security number or another identifying number or code; telephone number; or email address. PII can also include non-specific data elements such as gender, race, birth date, or geographic indicator 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, or procedures), 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 offer covered entities a simple 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 “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.