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 (through acquisition of Universal Patient Key) came along, private patient information was protected simply by removing it wholesale. 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 their healthcare data in one file (e.g. their hospital stay) with their data in another file (e.g. their pharmacy prescriptions after they were 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", a 44-character unique placeholder 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 Data Set A can be combined with Data Set B using the patient tokens to match one record to another without ever sharing the underlying patient information.

Datavant technology is delivered as a Java package for use in Windows or Linux environments, and 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. All participants are given their own software package to install, which is a simple and well-supported process that has been performed at provider sites, analytics companies, large data aggregators, and pharmaceutical manufacturers.

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: de-identification of the data set (including both removal of patient information as well as modifications of patient information), and the insertion of encrypted patient tokens.

Datavant technology is HIPAA-certified to create statistically de-identified data sets through a combination of removing personally-identifiable information (PII) like names, medical record numbers, etc., and modifying other values, such as turning 5-digit zip codes into 3-digit zip areas, or dates of birth into years of birth. The technology can also support more complicated derivations of data such as shifting dates of service or replacing detailed medical codes (e.g. the ICD10 code for Ebola with

more general medical codes.

As the Datavant 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, and as such, are consistently created from any data set where the PII is the same. Therefore, these 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. The tokens to be created can be based on any set of PII available in the data set, and can be customized during the configuration process. Multiple tokens are often created to facilitate the matching process, and can include fields such as social security numbers for deterministic matching, or it can include fields such as names and dates of birth, that when combined can create a unique token for probabilistic matching. Over several years of production implementations, Datavant's QA testing protocols have shown that Datavant technology generates reliable, repetitive tokens.

Configuration of the De-Identification Process

The Datavant de-identification engine is fully configurable to work with any data layout, and to work with any set of de-identification rules the data owner wishes to apply. This configuration process is a collaborative effort between Datavant and the data owner to ensure that the resulting de-identification process satisfies the data owner's regulatory and business compliance terms.

The configuration process 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 source's technical resources (e.g., defining pipe delimited database extracts that will serve as the inputs to the process). Within the input data set, Datavant and the data owner then identify all PHI data elements and their positions (including identification of any obtuse PHI elements; e.g., the location of a service address if that service is defined as "home service"). These PHI elements are what will be removed or modified by the Datavant technology, according to the rules defined in the software's "template" file.

The template file contains all of the de-identification rules to be applied when the technology is run by the data owner. Datavant can configure the de-identification engine to de-identify data sets to completely remove PHI elements (e.g. using the Safe Harbor method outlined in HIPAA regulations), but data owners who seek to preserve the analytical value of their data sets commonly choose rules that allow statistical de-identification (e.g. the Expert Determination method outlined in HIPAA regulations). Datavant has commonly used scrubbing rules in place to support the Expert Determination methodology, for names, dates, gender, patient and subscriber zip codes, as well as other typical PHI values. Table 1 describes the rules typically applied each HIPAA-designated PHI element under the Expert Determination methodology

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

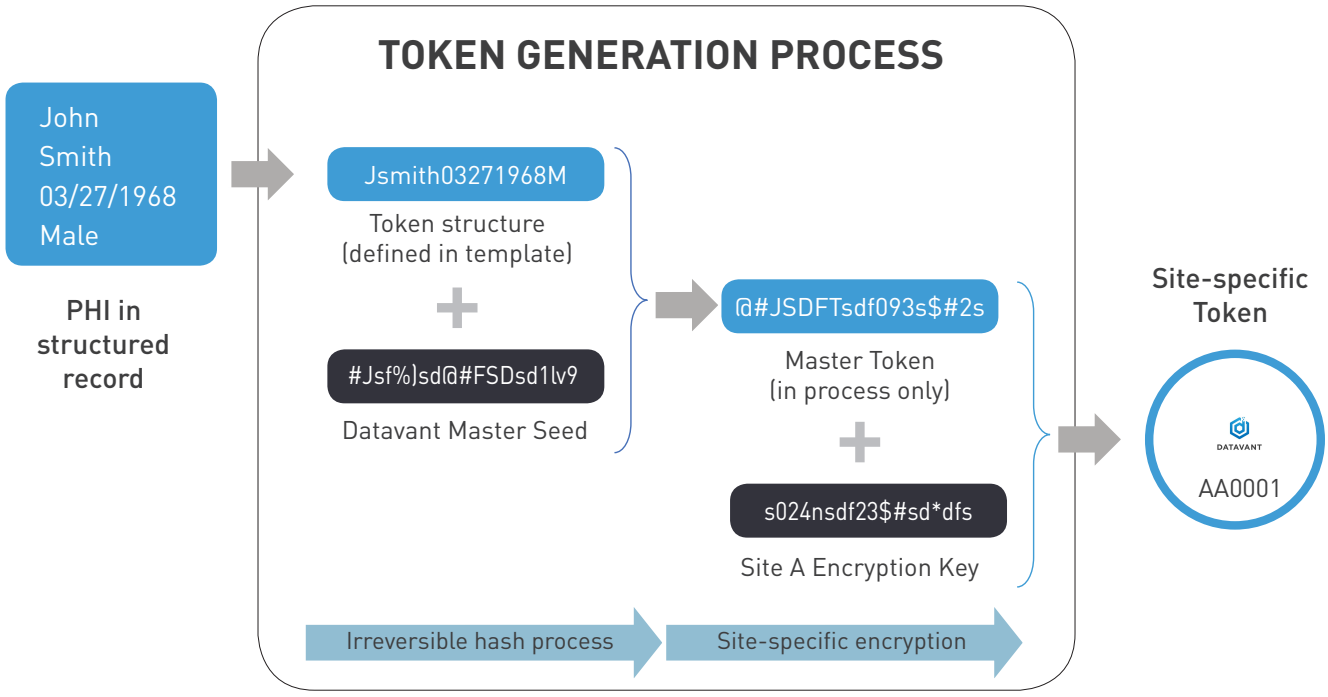
Names	Removed where present
Zip Code	All patient and subscriber zip codes are reduced to the initial three characters to define a zip area. Based on HIPAA rules, however, even three-digit zip areas with a combined population of 20,000 or less are either nulled out or are combined again with additional zip zones to ensure that populations exceed this minimum
Date of Service	Dates of service (e.g. admission dates, discharge dates, prescription fill dates, procedure dates, etc.) are typically preserved when using a statistical de-identification methodology
Date of Birth	The standard template's birthday rule converts all birth dates to January 1st of the birth year (the data source compliance officer may alter this to month versus year). All dates of birth where the individual would be 89 years of age or greater as of the date of the 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 file that is used to enforce the de-identification rules used by the Datavant technology is also used to define the PII values that are fed through the hashing and encryption process to create encrypted patient tokens. These tokens are used to identify and link matching individual records across different data sets without ever exposing the PHI of the patient to whom that record belongs (see explanation in the linking section later in this document).

Datavant has developed a unique method to ensure that encrypted tokens are (a) irreversible to reveal the patient's identity and (b) site-specific such that each user's tokens are unique from any other user's tokens (patent pending). It is critical for HIPAA compliance 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 that is destructive in nature, meaning that the patient's PII that is input into the token algorithm is unrecoverable from the output hash value. The second step of the token creation process is that the hash value ("Master Token") is encrypted with a site-specific encryption key to generate the final encrypted patient token. Thus, though the same patient information will always create the same Master Token, the site-specific encryption means that the Datavant de-identification engine will output a different token for that same patient in different users' data sets. In this way, a breach revealing a map of Datavant tokens and their originating PHI at one data site would never compromise the tokens or PHI at any other site using Datavant technology, as their tokens would not be the same even if they had the same patient in their data set.

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



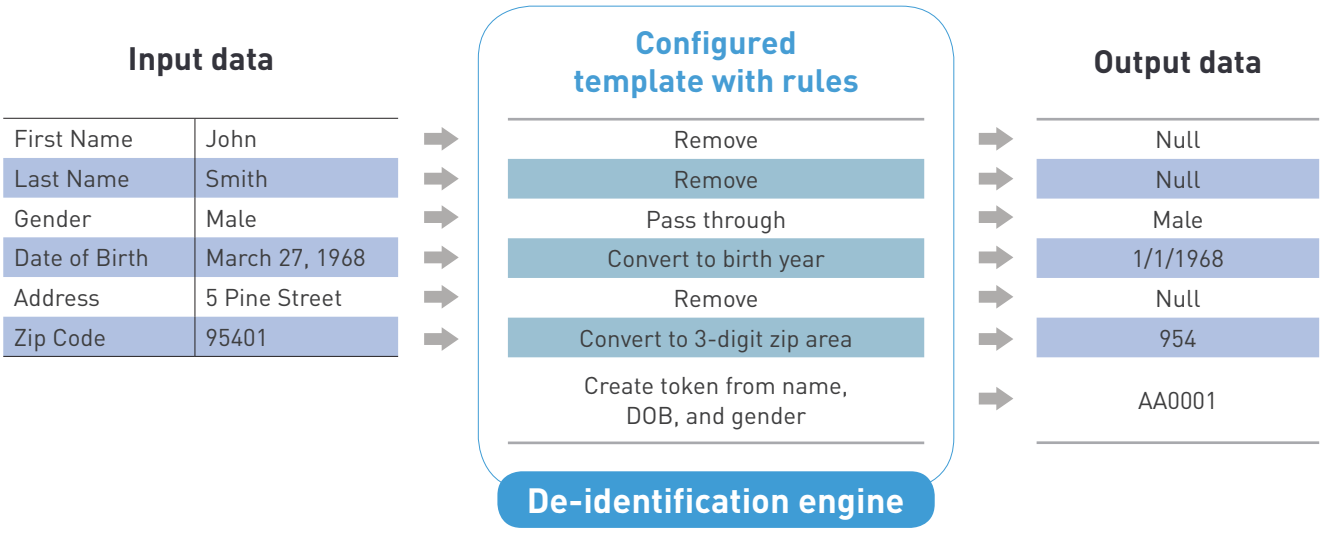
The PII elements that are to be used in the token creation process are defined in the template file. Datavant works with data owners to determine which fields should be used to create tokens, which is primarily the result of determining which PII fields are present in all of the different data sets the data owner wishes to join.

Datavant recommends that every template include the same base combination of name, date of birth, and gender to create a pair of core token values indicated as "Token 1" and "Token 2" in the template layouts, as these fields are present in almost every data set and generate match rates of 99% accuracy (see our Whitepaper titled, "Matching Accuracy of Patient Tokens in De-Identified Health Data Sets: A False Positive Analysis" on our website www.datavant.com). Additional token values can be generated for each record to facilitate matching across data sets, using unique values from the data stream that assist in linkage of data where full name, date of birth, or gender are not present consistently within the input file.

Using the Configured Template to Generate De-Identified Data Files

Once all of the de-identification and token creation rules are defined and mapped to the appropriate PHI element within the input data set, the final template is developed and embedded within the Datavant software package that the data owner uses to process each input data file (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: Datavant technology uses the template file to reliably and compliantly generate the de-identified and tokenized output file



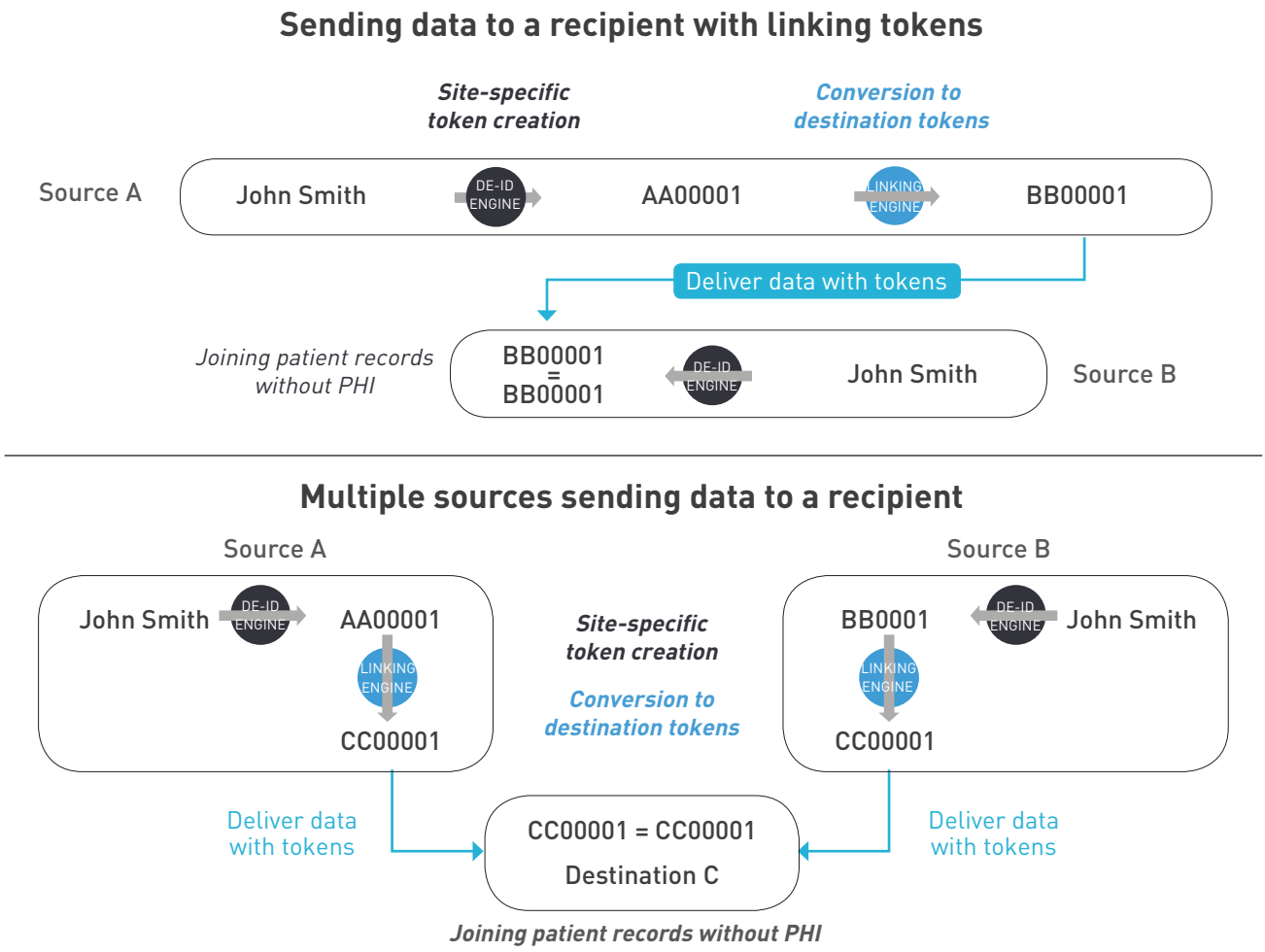
Token Transformation to Allow Matching Data Sets

As stated above, the output of the de-identification engine process creates de-identified data sets with site-specific tokens added to each patient's record. To merge disparate data sets, therefore, a second step is required whereby the site-specific tokens from each party are transformed into the same token scheme so that matching patient records can be identified. We provide a second module, focused on linking, to perform this conversion.

Like our structured data de-identification engine, our linking engine is a Java-based software package that is installed and run locally at the site. It allows the data source to convert the site-specific tokens they have created during the de-identification process into the token scheme of the data recipient. To link the data sets of each two different sites together, the tokens generated at the data source need to be converted to those of the data recipient. The linking engine performs this conversion task, ensuring that the tokens in the output files are now match-able.

Datavant recommends that all token conversion occurs at the source, such that each site is never sending its tokens on to another party, but is always sending data in the data recipient's token scheme. However, Datavant can perform token conversion at the data recipient's site if a central conversion process is desired. See Figure 3 below for an illustration of the token transformation process to allow matching of records.

Figure 3: Datavant tokens allow joining of a patient's records across data sets without ever sharing PHI



The primary feature of this token transformation process is that the patient key is never exposed. The site keys are unrelated to the data at the site and therefore, if site key token values were to become exposed, Datavant can provide new site keys to any of its clients along with new translation programs. These translation changes can even be applied to existing data sets, completely 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. Datavant's linking engine is only able to transform tokens from one site into another site's key with permission from that other site. Thus, data sources are able to control the link-ability of their data by granting or rescinding permission to convert the source's tokens into the recipient's tokens.

Flexible Usage of the Technology

Datavant's technology is a Java-based set of executables delivered as a JAR file to the environment where data is being processed, either at disparate data sources or a centralized data aggregators and data users. The software is installed and run at the site itself – no PII is required to be sent out from the site, Datavant has no access into the site's system.

Datavant's technology can be run to process files (batch mode) or to process individual strings of records (API mode). Both modes are available from the same software installation and are selected through the command line prompt. In this way, the software can be used to run a de-identification and tokenization routine on files as they appear in an input queue, or the software can be used in API mode to de-identify and tokenize records as they are being processed, loaded, transmitted, etc. as part of other routines.

And as discussed above, because the de-identification and tokenization rules are dictated by template files, the software can be easily configured to support any data layout, and to support any mixture of data modification or removal rules for effective and compliant de-identification. Further, users can store multiple template files, allowing different rules to be applied to a data set depending on which template is chosen by the user, and allowing one implementation of the software to process different data sets with different layouts.

How Our Clients Take Advantage of Datavant's Technology

Datavant is installed across the healthcare spectrum, as all stakeholders face the common issue of needing to protect patient privacy in their data sets while maximizing that data's utility in healthcare analytics. Datavant offers a simple, reliable, and flexible solution to de-identifying datasets in a HIPAA-compliant manner while still retaining the ability to link data sets from multiple sources without exposure of PHI. Below we list the distinct advantage we offer key stakeholders across the healthcare continuum:

Healthcare Providers – Clinicians and facilities providing direct patient care.

- Datavant offers healthcare providers the greatest degree of HIPAA compliance when managing their healthcare data beyond direct patient care. Datavant protects against regulatory violations by employees or Business Associates who do not have a legitimate need access to PHI.

Healthcare Payers – Organizations that pay for healthcare services.

- Datavant offers healthcare payers the simplest way to merge and analyze healthcare data while achieving the highest degree of HIPAA compliance.

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

- Datavant offers healthcare service providers the safest way to merge, share, or sell healthcare data to multiple organizations while maintaining the highest degree of HIPAA compliance.

Pharmaceutical and Medical Device Manufacturers, Data Aggregators and Analytics Firms – Companies that 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 is HIPAA-compliant, addresses institutional review board (IRB) requirements, and meets rigorous scientific standards.

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.