Executive Summary

The healthcare industry continues to work toward the goal of seamless health information exchange and data interoperability. With the vast amount of healthcare information available, the need to integrate data in real-time is a critical component of any solution aiming to produce actionable analytics that can improve patient care delivery, optimize supply chain management, or ensure accurate and timely reimbursement. However, the complexity of healthcare information, the data structure variation across key entities (e.g., providers, payers, and pharmacies), and stringent HIPAA privacy regulations all create barriers for seamless, real-time processing of healthcare data. This paper examines the challenges for de-identifying healthcare data as it is loaded into a data warehouse in real time, and details a solution to overcome these barriers in a reliable and automated fashion.

Key Findings Include:

- HIPAA regulations mandate that patient data is de-identified prior to use for non-clinical purposes like analytics, which has added time-consuming process barriers to real-time data integration efforts. When protected health information (PHI) is brought into a data warehouse, it is typically removed through a periodic batch processing of file extracts, which limits the potential for real-time analytics and interventions.

- Healthcare data is transmitted via Electronic Data Interchange (EDI), a computer-to-computer exchange of data in a standard electronic format between healthcare entities. Healthcare EDIs exist in many standard, structured formats. Common structures vary by health entity type (e.g., payer, pharmacy, or provider), but the standards ensure information integrity and reliability when sharing information within and between health entities. Common data formats include Health Level Seven (HL7) messages for clinical and administrative data, Accredited Standards Committee (ASC) X12 transaction sets for healthcare insurance claims, and the National Council for Prescription Drug Programs (NCPDP) transaction sets for pharmacy claims, among others.
• Datavant has developed technology that performs patient de-identification and tokenization in real time. Datavant accepts continuous incoming streams of data (as opposed to processing distinct files of data in a batch mode) in the native data form (e.g., HL7, X12 or NCPDP) via API and de-identifies and tokenizes the data before it is loaded into the final enterprise data warehouse. Real-time data de-identification removes a significant barrier to developing actionable, real-time analytics in a HIPAA-compliant manner.

• Using the Datavant API, Sentry Data Systems built a real-time de-identification workflow - the first of its kind - to bring together data from over 8,000 U.S. healthcare entities, representing 100 million patients, in a de-identified and linked logical data warehouse.

Introduction

Healthcare organizations are continuously generating new streams of data. These data contain a rich array of information that can be analyzed for many purposes, such as improving patient care, identifying adverse drug events, optimizing the supply chain, or developing tactics for increased brand adoption. While these healthcare data sets are full of information that can be used to answer a range of business questions, they are also full of PHI.

To tap into healthcare data for non-clinical use, data users must take extra steps under the security and privacy protections outlined 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. HIPAA specifies a set of PHI elements that must be removed to de-identify or “anonymize” a patient’s record in order to protect patient privacy. Such data de-identification can occur via the Safe Harbor method, in which all of the information that falls within eighteen HIPAA-specified categories (see Glossary) is removed; or by using the Statistical method, which retains more patient demographic information (but never enough PHI to re-identify any single patient). Both methods of de-identification create data sets that can be deemed adequate under HIPAA’s privacy and security rules.

The need for HIPAA-compliant data de-identification poses an additional challenge for organizations that may have access to a number of different data sources, but who want to create a single, longitudinal healthcare data set for analytics. Once two healthcare data sets are de-identified, patient records ordinarily cannot be accurately matched across the data sets because the identifying information necessary to link the records has
been removed. However, with Datavant’s technology, records are not only de-identified; each is also represented with a unique, encrypted token. The software consistently reproduces the same unique token for a given patient, regardless of the data set, and regardless of where the software is run. These unique tokens can then be used to match an individual’s records across de-identified healthcare data sets, providing users a method of merging data sets to create a single, longitudinal data set without ever seeing PHI.

The Rationale for Real-Time Data
An important feature of a healthcare data set is “data lag”, which can be measured as the time between when an analyst receives the data and the last dates present in the record set. Data lag is primarily determined by the frequency of data updates. While some healthcare analytics can rely on infrequently updated (e.g., monthly or quarterly) static data sets, other applications require access to real-time data. For example, static, dated data may be sufficient for answering high-level strategic questions, such as determining the outcomes of a particular treatment paradigm, or assessing the prevalence of a disease for drug forecasting or hospital planning. By contrast, a pharmaceutical company’s brand team relies on access to frequently updated utilization data (e.g., prescription new starts, refills, and switches) to track brand performance, and to make rapid tactical adjustments. Similarly, health providers and payers require real-time data to implement population health programs, patient support programs, and safety programs like adverse event alerts. Real-time data is also critical for optimizing supply chain management to ensure regional supply meets demand or for Risk Evaluation & Mitigation Strategies ("REMS") reporting and interventions so that any emergent patterns of adverse events can be addressed immediately.

Real-Time De-Identification Streamlines Data Processing for Real-Time Analytics
Data de-identification often occurs via a batch process (Figure 1A). In this case, healthcare data is sent to a data warehouse with PHI included. To remove the PHI, a file extract of the database is created (a batch), which is run through de-identification and tokenization software, and then the de-identified and tokenized data is loaded back into the final data warehouse. Each time new data arrives, this batch process needs to be repeated, preventing users from accessing data in real-time. If de-identification and tokenization occurred as each string of data entered the data warehouse, then PHI would never be stored in the data warehouse, and users could continuously run real-time analytics in a HIPAA-compliant manner.

To support real-time de-identification and tokenization, Datavant employs an application programming interface (API), a software-to-software intermediary that makes it possible for two applications to seamlessly
talk to each other and securely share data. Datavant’s software can accept continuous incoming streams of data (as opposed to distinct files of data) via API, and de-identify and tokenize the data before it is loaded into the final data warehouse (Figure 1B). With this process, PHI is never stored in the data warehouse, and the laborious and cumbersome batch process is rendered unnecessary.

**Figure 1: Batch Versus Real-Time (API) De-identification Process**

**Real-Time Processing from Native Data Formats**

As illustrated in Figure 1, while the batch de-identification process occurs after a standardized data extract process, with Datavant’s technology, the API is accessed before any data is loaded into the database. As a result, Datavant’s technology must be able to process the incoming healthcare data in its native form.

Healthcare data is transmitted via EDI, a computer-to-computer exchange of data in a standard electronic format between businesses, such as two healthcare entities (e.g., health plans, health clearinghouses, or providers). The standard language, structure, and content of an EDI transaction allows two different healthcare entities, such as a provider and a payer, to seamlessly exchange information electronically. For example, the standard EDI X12-837 transaction set enables providers and payers to communicate about patient insurance claims. The common format of the 837 transaction allows providers to submit claims and receive payment information with any payer regardless of the specific IT platform each party uses. These EDI
transactions support organized and efficient healthcare information exchange, and also make such data sets valuable candidates for data analytics.

The standardization of EDI formats supports interoperability between disparate health systems and health entities, and consistent data mapping for use in downstream analytical applications. Standards developing organizations (SDOs) develop and maintain electronic healthcare data formats. For example, the American Society for Testing and Materials (ASTM) and Health Level Seven (HL7) develop clinical data standards, the Accredited Standards Committee (ASC) X12 oversees insurance and remittance standards, the Digital Imaging and Communications in Medicine (DICOM) develops diagnostic image standards, and the National Council for Prescription Drug Programs (NCPDP) oversees pharmacy transaction standards. A selection of common healthcare data standards and structures that are relevant for real-time data analytics are shown in Table 1.

<table>
<thead>
<tr>
<th>Healthcare Data Format</th>
<th>Select Transaction / Message Sets</th>
<th>Type of Healthcare Data</th>
<th>Healthcare Entities Who Exchange Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC X12</td>
<td>837</td>
<td>Healthcare Claims (Institutional, Professional, and Dental)</td>
<td>Provider : Payer</td>
</tr>
<tr>
<td></td>
<td>835</td>
<td>Healthcare Claim Payment</td>
<td>Payer : Provider</td>
</tr>
<tr>
<td></td>
<td>832</td>
<td>Drug Price / Sales Catalog</td>
<td>Manufacturer : Supply Chain</td>
</tr>
<tr>
<td>NCPDP</td>
<td>D.0</td>
<td>Retail Pharmacy Claims</td>
<td>Pharmacy : Payer</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
<td>Medicaid Pharmacy Subrogation</td>
<td>Pharmacy : Payer</td>
</tr>
<tr>
<td></td>
<td>SCRIPT</td>
<td>ePrescribing</td>
<td>Provider : Pharmacy : Payer</td>
</tr>
<tr>
<td></td>
<td>Formulary and Benefit Standard</td>
<td>Patient and Drug- Specific Insurance Information</td>
<td>Provider : Pharmacy : Payer</td>
</tr>
<tr>
<td>HL7</td>
<td>HL7 Version 2 Product Suite</td>
<td>Clinical Messaging Standard</td>
<td>Provider : Provider</td>
</tr>
<tr>
<td></td>
<td>Version 3 Product Suite</td>
<td>Clinical Messaging Standard in XML Syntax</td>
<td>Provider : Provider</td>
</tr>
<tr>
<td></td>
<td>Clinical Document Architecture (CDA) Release 2 and Consolidated CDA Documents (e.g., CCD*)</td>
<td>HL7 V3 Clinical Document Standards</td>
<td>Provider : Provider Provider : Patients Provider : Payer : Pharmacy</td>
</tr>
<tr>
<td></td>
<td>Fast Healthcare Interoperability Resources (FHIR)</td>
<td>Next Generation Clinical Standards</td>
<td>Provider : Provider Provider : Payer Provider : Patient</td>
</tr>
</tbody>
</table>

Note: “CCD (Continuity of Care Document) was a joint development effort between HL7 and ASTM. The CCD is based on the content in ASTM's CCR (Continuity of Care Record), but built under the architecture of HL7's CDA. U.S. regulations note the CCD as the preferred format of electronic clinical record exchange.

Electronic Health Records are a rich source of patient treatment histories. Widely implemented EHR standards are not yet available, although multiple organizations are working toward solutions. HL7's emerging standard, Fast Healthcare Interoperability Resources (FHIR), which builds on HL7's v2, v3, and CDA standards, is currently published in the Standard for Trial Use stage, and has the potential to greatly improve electronic sharing of EHR data.
Because EDI messages are intended for computer-to-computer exchange of information, their visual format can be difficult to understand. An overview of the hierarchical structure of an example X12-837 transaction set, and a sample of a message segment is shown in Figure 2. In this example, the data transfer is called a transaction set. The transaction set is broken up into standard sections (“Loops”), and each section provides a specific type of information. For an 837 transaction, the sections include a transaction header, billing provider detail, subscriber detail, patient detail, claim detail, and a transaction trailer (Figure 2a). In Figure 2b, a single section of the transaction set data code is displayed. Each section is further broken up into data segments, labeled with a combination of two or three alphabetical characters that begin a line of text (bolded in Figure 2b). Finally, each data segment is comprised of one or more data elements, which are delimited by a data delimiter (an asterisk, dash, or colon are commonly used). While the data in EDI transactions are difficult to parse visually, the standard structure, content, and data format support reliable, accurate, and efficient healthcare information exchange.

Figure 2: Example of An EDI X12-837 Transaction Set Structure

ASC X12 Data Formats

ASC X12 is a consensus-based, American National Standards Institute (ANSI)-accredited organization that focuses on the development, implementation, and ongoing use of interoperable EDI standards. Membership spans multiple industries, with standards developed for healthcare, insurance, transportation, finance, and supply chain, among others. As part of the HIPAA-standardization effort for electronic health transactions, the Department of Health and Human Services (HHS) developed HIPAA-EDI transaction sets using the ASC X12 and NCPDP formats, both of which support the widespread use of EDI for health information exchange.

The X12 standards that relate to healthcare operations include transaction sets for insurance claims and claims-related processing (e.g., 837 claims and 835 claims payment), healthcare administration (e.g., 270/271 benefits eligibility and 834 plan enrollment, etc.), and healthcare procurement (e.g., 832 sales and pricing catalogs).

The X12 837 transaction set provides a means for providers to communicate with payers regarding claims processing and payment. The X12 837 data, which covers institutional, professional, and dental claims, provides analysts a rich source of information for patient treatment history. The 837 transaction set does not capture retail pharmacy claims, which are captured via the NCPDP transaction sets (see next section). The X12 835 transaction set is complementary to the 837 set, and is used by payers to communicate claims payment or advice back to providers. For example, an insurer can use the 835 standard to make a claims payment and/or send an Explanation of Benefits (EOB) remittance advice to a provider, either directly or through a financial institution.

The X12 832 price/sales catalog transaction set is used to request or provide product and service pricing, and is how manufacturers provide detailed product information and prices to their supply chain. The 832 data set may include manufacturer/contact information, terms of sale (including discounts), product ID and description, packaging, pricing, quantity and unit of measure. The 832 data set is an important source of information for determining hospital or pharmacy operating profitability and supply chain management.

NCPDP Data Formats

The National Council for Prescription Drug Programs (NCPDP) is an ANSI-accredited standards development organization that uses a consensus-building process to create national standards for real-time, electronic exchange of healthcare information in the pharmacy sector. NCPDP develops electronic standards for
prescribing, dispensing, monitoring, managing, and paying for medications. NCPDP’s HIPAA-EDI standard transaction sets include the telecommunication standard D.0 for retail pharmacy claims and the standard 3.0 for Medicaid pharmacy subrogation (Table 1). The NCPDP D.0 and messaging standards provide a structured means for pharmacies to communicate with payers for claims processing and payment.

NCPDP develops a number of additional EDI standards for the pharmacy industry, including standards for defining pharmaceutical billing units, handling prior authorization requests, ePrescribing, and submitting manufacturer rebate information. For example, NCPDP’s SCRIPT standard is widely used for the transfer of prescription data between pharmacies, prescribers, intermediaries, facilities, and payers. SCRIPT messages include a rich set of pharmacy data, including data for prescription information (e.g., new prescriptions, changes, refills, fill notifications, cancellations, or HWF), patient medication history, allergies, standardized instructions, and prior authorization, among other information.

NCPDP also maintains the Formulary and Benefit Standard, which provides payers a way to communicate formulary and benefit information to providers in real-time. This standard data transfer provides physicians information about a patient’s insurance coverage (e.g., formulary, step therapy, payer-specified alternatives, and cost to patient) during the prescribing process to support the physician’s treatment decisions.

**HL7 Data Formats**

Health Level Seven (HL7) is another important ANSI-accredited standards-developing organization working within the healthcare space, and specializes in clinical and administrative data. HL7 is most widely known for its messaging and document standards, which enable both internal and external healthcare applications and entities to exchange clinical and administrative data.

Within the United States, over 95% of healthcare organizations rely on the HL7 Version 2.x messaging standard for the exchange of a patient’s electronic clinical data between provider systems. There are numerous HL7 message types spanning a patient’s care, but one of the most commonly used is the Patient Administration or Admit, Discharge, and Transfer (ADT) message. This message supports a number of the central administrative functions in healthcare and includes patient demographic and visit information, triggers patient registration in the healthcare setting, and is used widely across encounter management.

HL7 published the Version 3 (V3) product suite, a clinical messaging and documents standards system
designed for semantic interoperability. HL7 V3 was designed to replace HL7 V2. However, a lack of backwards compatibility with the V2 messaging system, among other challenges, has resulted in minimal use in the United States. The Clinical Document Architecture (CDA) Release 2 and Consolidated CDA Documents are the HL7 V3 document standards. The Continuity of Care Document (CCD) (a type of CDA document) allows providers to electronically transfer critical elements of a patient’s health history as a patient moves from one treatment setting to another to “continue care” and has been recognized in U.S. regulations as a means to exchange clinical information.\(^\text{12}\)

**Additional Data Streams Undergoing Standardization**

While the healthcare industry continues to develop structured forms of data collection, a number of valuable data sets exist in either variable structure forms (e.g., vary by vendor) or in unstructured forms. Electronic health records (EHR) are a critical source of patient treatment data, demographics, and history, but EHR data structures vary by vendor or application. While current EHR standards are not yet widely implemented, the emerging HL7 FHIR standard has the potential to greatly improve EHR interoperability.

Automated medication dispensing (AMD) is another example of healthcare data with a non-standard structure. AMD is a decentralized pharmacy distribution practice whereby a remote machine stores, tracks, and dispenses medication at the point-of-care, typically within a hospital or long-term care facility.\(^\text{13}\) The data formats of AMD systems vary by manufacturer, but the data captured provide analysts with detailed drug utilization information in the in-patient care setting. Many AMD developers are moving toward developing interoperability solutions by integrating their AMD data with EHR systems, but data integration remains a challenge.

**Real-Time Data De-Identification and Integration Improves Efficiency of Real-Time Analytics**

The development of EDI standards has greatly improved the reliability and accuracy of healthcare information exchange, but HIPAA regulations have added time-consuming process barriers to real-time data integration efforts (see Figure 1). To overcome some of these hurdles, Datavant has developed technology that performs de-identification and tokenization in real time and can be used to support HIPAA-compliant analytics. Datavant uses an API to accept continuous streams of data directly into its de-identification and tokenization software, which supports the real-time flow of de-identified data into a data warehouse (Figure 3). Data users
no longer need to use time-intensive batch processing to remove PHI from data warehouses before performing analytics—the data arrive de-identified and tokenized, and ready for data integration.

**Use Case: Improving Care by Reducing Care Redundancy**

Improving healthcare information exchange and interoperability within provider care settings can have a meaningful impact on quality of care. Patients who seek treatment in different care settings, or in different care networks, run the risk of redundant treatments or the increased risk of adverse events resulting from uncoordinated treatments. Having real-time provider tools that integrate patient data across care delivery settings and networks can offer treatment providers a decision-support tool that increases treatment transparency, and can function to improve patient care, reduce treatment costs, and improve patient safety.

**Use Case: Value-Based Treatment Tools**

Seamless integration of real-time provider, supply chain, and payer data supports the development of value-based treatment tools. As hospitals continue to look for ways to control costs, and as care shifts to a value-based model, physicians will increasingly need tools to help them assess treatment value, which considers not only treatment cost but also treatment outcomes. By combining real-time supply chain data (e.g., product cost and information), formulary data (e.g., patient-specific plan/cost), and clinical outcomes measures (e.g., length of stay, rate of infections, patient satisfaction), providers could have a powerful decision-making tool to compare treatment options that is supported by real-time data.
Case Study: Implementing Real-Time Data De-Identification and Tokenization Across a Network of over 8000 Providers

Datavant offers the de-identification and tokenization tools to support HIPAA-compliant, real-time analytics. Datavant's API includes configurations for HL7-ADT messages, X12-837 and 832 transactions sets, NCPDP transaction sets, and Automated Medicine Dispensing (AMD) data formats, which were used in implementing an integrated solution for a leading healthcare analytics company.

Sentry Data Systems is leading the healthcare industry in utilizing advanced data analytics to reduce the total cost of care, improve quality, and improve patient outcomes. More than 8,000 hospitals, clinics, integrated delivery networks and pharmacies across the country rely on the company’s unique, proprietary data set for analytics, procurement, drug utilization, and compliance solutions.

By implementing a real-time solution, the first of its kind, Sentry hopes to unlock a wealth of data available for over 100 million patients across the United States through a state-of-the-art logical data warehouse (LDW). The data to be contained in this warehouse consists of over 1,000 unique fields, including registration, diagnoses, procedures, pharmacy, lab and radiology orders. The warehouse also needed to include insurance claims and prescription information, and link all of these records together to allow analyses of a patient’s entire treatment experience. And it all needed to be de-identified.

Working in partnership with a large data integrator, Datavant worked to embed its API-enabled de-identification and tokenization software at the data ingestion point for Sentry. The software was configured to parse the complex data feeds streaming in from thousands of providers, pharmacies and clinics across the United States, including data transmitted in HL7-ADT, 837, 832, NCPDP, and AMD formats. These data feeds are de-identified on the fly according to the configured rules in Datavant’s software, and unique encrypted tokens are added back to each record to allow the linking of matching patient records downstream. The outflowing de-identified data is then fed into a state-of-the-art LDW that was built with a three-layer architecture:

1. Big Data Lake built on Cloudera Hadoop distribution
2. Operational data store built on MongoDB, integrating data from the 20+ feeds across 8,000+ U.S. healthcare entities
3. Data warehouse to provide standardized analytics and reporting capabilities
This real-time solution gives Sentry a robust, scalable, and secure platform Sentry can use to support an even broader spectrum of pharmaceutical, payer and provider use cases in a HIPAA-compliant way.

Sources

For More Information:

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

Connecting the World’s Health Data

Datavant helps organizations safely protect, link and exchange 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 exchange.

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, 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.