DATAVANT

Mortality Data in Healthcare Analytics

Sourcing Robust Data in a HIPAA-Compliant Manner

Executive Summary

Incorporating mortality information into healthcare data can strengthen identity protection, reduce healthcare costs, and improve health treatments and care delivery. Today, organizations use mortality data to help prevent fraud, perform accurate billing and benefits distribution, and to conduct outcomes analysis (particularly in areas like oncology, where survival is a key endpoint).

The vast majority of healthcare data sets do not capture mortality data. Historically, the Social Security Administration's Death Master File (DMF) has been a leading public data source for mortality records. However, recent changes to the DMF have dramatically reduced coverage from 2.5 million lives in 2010 to only 460,000 lives in 2016.

In this paper, we discuss options for filling the gaps in DMF data coverage, and present a solution for joining this enhanced mortality data to healthcare data sets in a manner compatible with the Health Insurance Portability and Accountability Act of 1996 (HIPAA).

Key findings include:

- Combining obituary data with the DMF is a robust method to increase U.S. mortality data coverage; in 2016, obituary data added 1.7 million unique records not present in the DMF data
- Transferring raw mortality data to a healthcare data environment violates HIPAA, as mortality data contains a number of fields that constitute personally identifiable information (PII; including but not limited to protected health information, or PHI)
- Prior to being transferred to a healthcare data environment, mortality data should be de-identified in accordance with HIPAA
- Using Datavant's technology to generate encrypted tokens allows mortality records to be linked to other health data records without reference to the underlying PII

The Value of Mortality Data in Healthcare Analyses

Over three decades ago, a United Nations report noted that “the sector in which mortality information is most directly valuable is that of health.”[1] Providers use mortality data to identify high-risk patient groups to inform resource allocation and improve health outcomes. Payers use mortality data to protect patient identity, limit fraud, and ensure accurate billing processes. In clinical development, mortality may be an endpoint or otherwise inform outcomes analysis. In short, mortality data can help reduce costs and improve the quality and outcomes of patient care.

Unfortunately, most healthcare data sets do not contain mortality data. Healthcare data is often gathered as part of an interaction with a site of care: physician notes entered into an Electronic Medical Record (EMR), an insurance claim filed after a laboratory test, a prescription filled at a pharmacy. Unless a patient dies in a hospital bed, mortality data will be missing from these data sets because deceased patients do not interact with the healthcare system.

Mortality Sources: The Social Security Death Master File

Historically, the DMF has been one of the main data sources healthcare organizations look to in order to fill mortality data gaps. The DMF is an index of over 85 million records based on SSA payment records and death reports from family members, funeral homes, financial institutions, postal authorities, states, and other federal agencies, from 1936 to the present.[2,3,4] Each DMF record typically includes social security number, full name, date of birth, and date of death. The file is updated weekly, making it a highly valuable and timely record of U.S. deaths.

While the SSA does not guarantee data completeness or accuracy, the DMF has become an important data source for death verification and fraud protection. Medical researchers rely on the information to track study patients and verify death, while members of financial institutions, insurance companies, and governments rely on the information to verify identity and prevent fraud. The National Death Index (NDI), maintained by the Centers for Disease Control (CDC), is a more complete record of U.S. deaths, but the DMF is more affordable, easier to use, and updated more frequently.

Under the Freedom of Information Act (FOIA), the DMF has been made available as a public file since 1980.

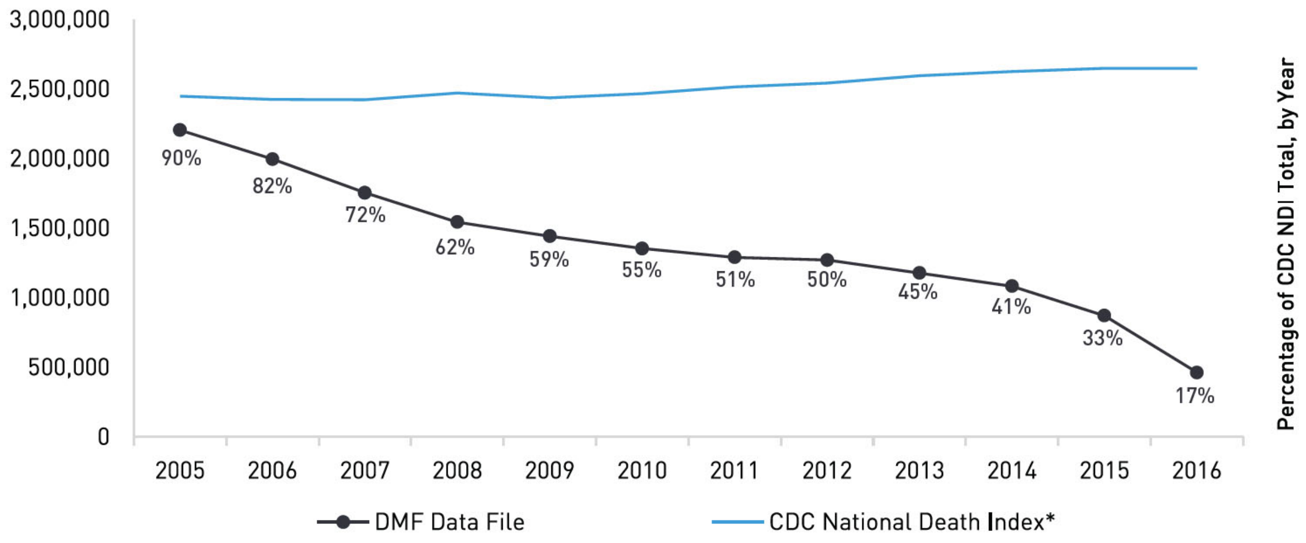
However, access to the dataset is contingent upon both an application and an annual certification.[4] The SSA grants full access to approved state and federal agencies, while the Department of Commerce's National Technical Information Service (NTIS) sells access to approved private and public organizations.

The DMF Public File: A Diminishing Resource for Mortality Data

In 2011, the usefulness of the DMF as a comprehensive record of U.S. deaths changed dramatically when the SSA stopped releasing state-level records. For a decade, the DMF included state-reported deaths, but amid rising concerns that the file provided identity thieves easy access to PII, the SSA determined it had been erroneously disclosing state records. The same year SSA removed 4.2 million historical death records and stopped releasing state-reported death records in subsequent updates.[5] In 2010, the SSA DMF file (with state records) included 2.5 million records[6]; by 2016, the file included only 460,000 records.

The impact of the SSA's decision on the completeness of the DMF is illustrated in the figure below. In 2005, the DMF accounted for 90% of the deaths reported by the CDC; in 2016, the DMF accounted for a mere 17% of CDC-reported deaths. For healthcare data analysts, the sudden lack of reliable, timely mortality data could easily lead to billing and benefits errors, slow clinical trial and long-term epidemiological studies, and hamper outcomes analyses.

Figure 1: Annual U.S. Deaths (2005 – 2016): Reported Volume in DMF Data File vs. CDC NDI



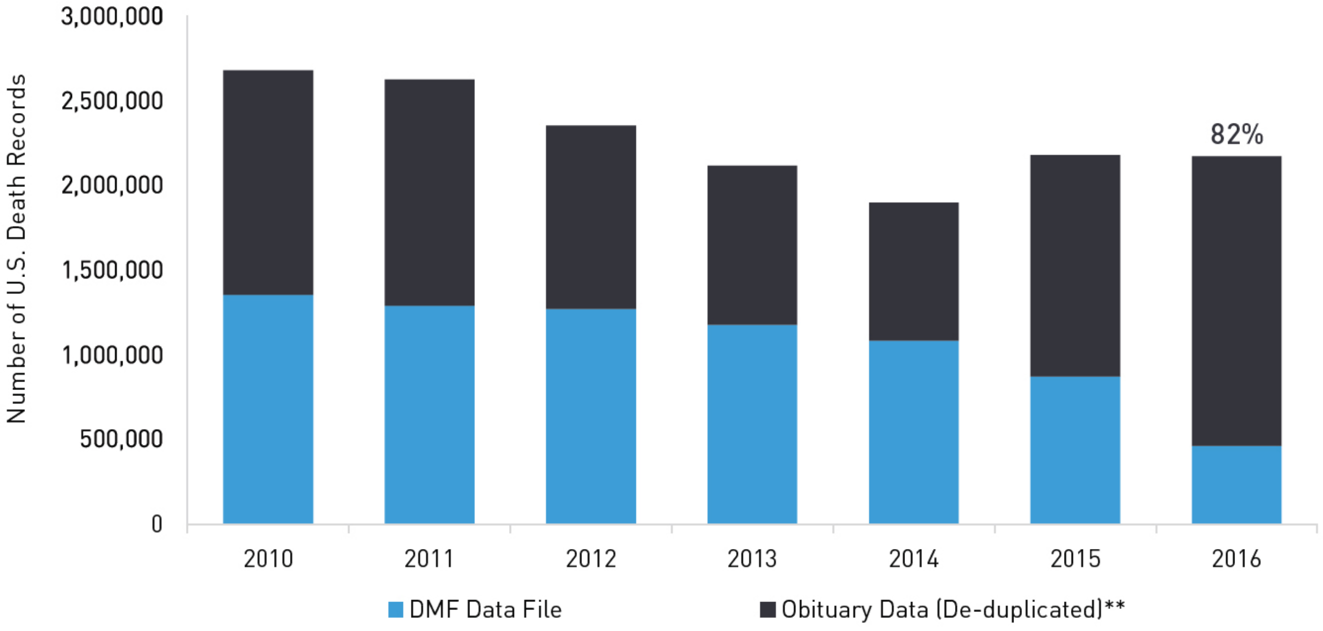
*Note: *CDC NDI values for 2015 and 2016 are estimates.*

Leveraging U.S. Obituary Data to Increase Mortality Coverage

Given the continued decline in DMF coverage, we wanted to understand the extent to which obituary data might increase mortality coverage beyond that of the DMF alone. To do so, we joined obituary data gathered since 2010 to the DMF, removed duplicate individuals shared between the two files, and counted the total number of unique records per year. We used the CDC's NDI data as a benchmark for total U.S. deaths in a given year.

As shown in Figure 2, adding obituary data to the DMF significantly increases mortality data coverage. In 2016, the inclusion of obituary data added 1.7 million unique mortality records to the 460,000 included in the DMF for that year. The combined data, shown in Figure 2, contains nearly 2.2 million individual mortality records for 2016, or 82% of those captured in the CDC NDI. Augmenting DMF data with obituary data is an effective and robust method of capturing the majority of U.S. mortality records.

Figure 2: Filling the DMF Data Gap with Obituary Data



Notes:
**Percentage shown is the percentage of the total estimated CDC NDI death records in 2016 represented by the combined mortality data set.*
***Mortality records present in both the DMF data file and in the Obituary records were removed from the Obituary data.*

Incorporating Mortality Data in Healthcare Environments: HIPAA Considerations

Healthcare organizations must abide by the security and privacy regulations 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 of 2009. In accordance with government regulations, records must be de-identified before being incorporated into any other healthcare data set that is intended to be used for anything other than direct clinical support.

Incorporating mortality data into existing healthcare data sets presents security and HIPAA-compliance challenges. Publicly and privately-available mortality data includes a wealth of PHI and PII (e.g., names, addresses, dates of birth and death, and social security numbers). Therefore, adding mortality data directly into a healthcare data set could be a HIPAA violation.

To ensure HIPAA-compliance, mortality data should be de-identified prior to being transferred to a healthcare environment. However, a standard de-identification process presents an obvious challenge. After removing all PHI and PII, data users have no way to match deceased individuals in mortality data sets to de-identified individuals in other healthcare data sets.

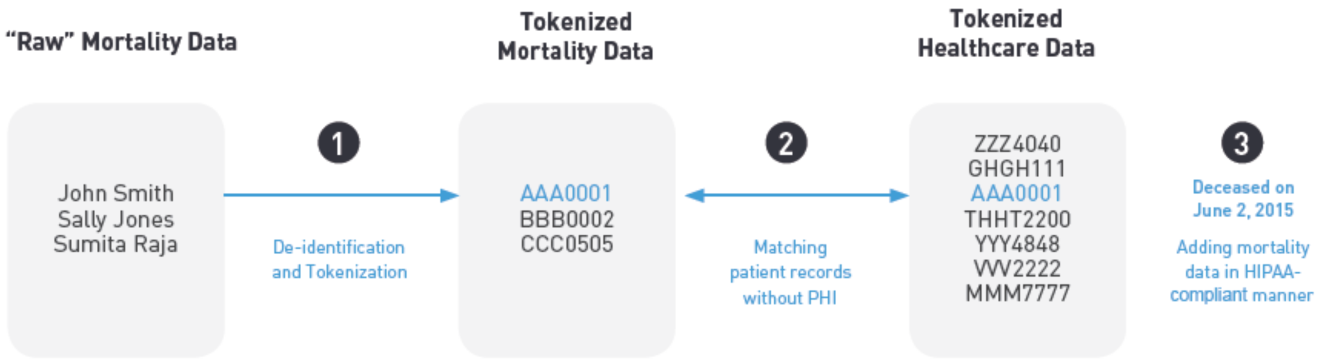
De-Identification and Linking of Mortality Data Using Patient Tokens

To ensure our combined mortality data set could be deemed de-identified under the HIPAA Expert Determination method, we processed the death records using Datavant's technology. To statistically de-identify the data set, the technology takes in the raw mortality data, and removes names, addresses, and social security numbers; as well as making other modifications like changing date of birth to year of birth.

In the process of de-identification, Datavant's technology adds a unique, encrypted token to each record, which is built from the underlying personally identifiable information. These tokens correspond to unique individuals, and so can be used to link mortality data to other healthcare data sets.

Clients may link de-identified, tokenized mortality data to any similarly de-identified and tokenized healthcare data set in a HIPAA-compliant manner. Without ever seeing or holding the underlying PII or PHI, healthcare analysts can simply match tokens across the different data sets to determine which patients should be identified as deceased (Figure 3).

Figure 3: Linking of Mortality and Healthcare Data Sets: Patient De-identification, Token Creation, and Linking of Data Sets



Patient records in the healthcare data set(s) can now be updated with death information, while minimizing the risk of HIPAA violations by ensuring that users are never exposed to PHI.

For a detailed description of our de-identification and linking process, please see our whitepaper:

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

Characterizing Tokens to Enhance Data Certainty

For healthcare analysts interested in incorporating mortality data into other healthcare data sets, there are two key expectations: (i) that patients in the mortality data set are actually deceased, and (ii) that the de-identified deceased patients are accurately matched across healthcare data sets.

To ensure that these expectations are met, Datavant developed a death validity score and a uniqueness score for each record. The death validity score indicates our degree of confidence that an individual in a given mortality data set is actually deceased. The uniqueness score indicates the likelihood that an encrypted token for a given mortality record designates a unique person (i.e., that there are not two distinct individuals for whom Datavant's technology generated the same token).

The Death Validity Score

As we noted above, the death validity score indicates our degree of confidence that an individual in a given mortality data set is actually deceased. It is possible to reflect the death validity score as a quantitative value, but given the depth of existing mortality data, it is our view that assigning a specific numeric value would reflect false precision.

To assign a death validity score, we apply a set of simple checks. The first is a check against the DMF to see if the record has either a "Proof" flag (meaning the SSA received a death certificate) or a "Verified" flag (meaning a family member verified the death). The second check examines the remaining un-flagged records to see if specific individuals are identified as deceased in both the DMF and obituary data files. A record that passes the first check has the highest death validity score; a record that passes the second has an intermediate death validity score; and a record that passes neither check has the lowest death validity score.

The Uniqueness Score

The uniqueness score indicates the likelihood that an encrypted token for a given mortality record designates a unique person. As discussed in our white paper, *Matching Accuracy of Patient Tokens in De-Identified Health Data Sets: A False Positive Analysis*, because tokens are built from underlying elements of PII, it is possible that two distinct individuals may share the same token. For example, if a token is built only from name, gender and birthday, then two men born on the same day and both named John Smith would share the same token. Our study indicates that for our most commonly used tokens, non-uniqueness occurs in only about 1% of cases.

If a token is unique in the combined mortality data set, we assign a score of 100%. Alternately, if a token is shared by four different individuals in the mortality data set, we assign a score of 25%, meaning matching this token against a healthcare data set has only a 1 in 4 chance of being accurate.

Study Conclusions

Based on this study, we have reached a handful of conclusions:

1. Combining obituary data with the SSA DMF creates a much more robust and representative mortality data set than using either source alone.
2. By de-identifying and tokenizing the mortality data set with Datavant's technology, we can link mortality data to other healthcare data sets without any exchange of PHI, thereby minimizing the

risk of a HIPAA violation.

3. The Death Validity score allows healthcare analysts to assign a qualitative score to death records based on the likelihood that the person represented in the record is actually deceased.
4. When linking mortality data and other healthcare data sets, our most frequently used tokens have a match rate of 99%, and we can identify the 1% of records for whom a match may be a false positive.

Datavant's unique ability to provide robust mortality data that can be linked to other healthcare data sets in a HIPAA-compliant manner is being used by a number of customers today to prevent identity theft and fraud, improve interactions with patient communities, and to perform valuable population health and outcomes analytics.

Sources

1. *www.ncbi.nlm.nih.gov/nlmcatalog/101021439*
2. *www.ssdmf.com*
3. *https://www.ssa.gov/dataexchange/request_dmf.html*
4. *https://classic.ntis.gov/products/ssa-dmf/#*
5. *http://oig.ssa.gov/newsroom/congressional-testimony/hearing-social-securitys-death-records*
6. *http://www.nytimes.com/2012/10/09/us/social-security-death-record-limits-hinder-researchers.html?_r=1&*

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.