Veeva Crossix

NOVEMBER 2019

Beyond HIPAA:

Risks of Re-identification in Today's Digital-Centric World

Abstract

In the healthcare industry, the ability to combine health data with other sources of data, such as consumer and media data, is a powerful new tool allowing pharmaceutical companies to uncover new insights into patient behavior and make more informed business decisions. In particular, marketers have the ability to connect health data with digital media data, including information about ads seen or websites visited by an individual. However, combining these disparate data sets can raise serious privacy concerns. Existing regulatory frameworks may not go far enough to fully protect privacy in this environment on their own, but certain self-regulatory efforts go a long way toward closing the gaps.

There are many analytics approaches that use anonymized datasets to combine health and digital data. As the opportunities grow for companies to combine ever-expanding sets of data, the potential for re-identifying an individual's personal medical records increases as well.

This whitepaper will provide a summary of regulatory and industry guidance and an overview of current analytics approaches, along with the relative re-identification risk of each one.

Key Takeaways

- The Health Insurance Portability and Accountability Act (HIPAA) is not the only relevant regulatory framework covering the use of de-identified data, including in light of efforts at the state level. On the self-regulatory side, the Network Advertising Initiative (NAI) provides a different and equally important way to evaluate approaches to digital media measurement.
- When using de-identified data, companies must preemptively evaluate the risk of re-identification against regulatory thresholds, industry norms and best practices for protecting patient privacy.
- Simply getting a HIPAA statistical certification that data is appropriately de-identified may not be enough to protect privacy. Metadata present within a health data set or when combined with other data sets can easily allow for the re-identification of an individual's personal health information.
- Privacy leaders should consider whether compliance with a regulatory framework is accomplished via agreements/trust or via technology. Only the latter truly prevents re-identification of health data.
- To be compliant with NAI guidelines for Ad Delivery and Reporting, an approach should not enable the re-identification of Personally-Identified Information (PII) or Device-Identified Information (DII) in combination with healthcare data. Keeping data in separate locations and removing personal information is not sufficient to protect privacy.

Legal and Regulatory Guidance

The Health Insurance Portability and Accountability Act (HIPAA)

Enormous value can be created in healthcare by combining traditionally siloed claims, clinical and health data with information that completes the picture of the patient as an individual. "The de-identification of protected health information (PHI) enables HIPAA covered entities to share health data for large-scale medical research studies, policy assessments, comparative effectiveness studies, and other studies and assessments without violating the privacy of patients or requiring authorizations to be obtained from each patient prior to data being disclosed."[1]

HIPAA provides a framework for companies working with de-identified data, and the law recognizes the risk that combining different types of data could enable re-identification of an individual. Under the Expert Determination method for de-identification, an expert statistician must approve each use case for connecting protected, de-identified health data. As part of that evaluation, the statistician must:

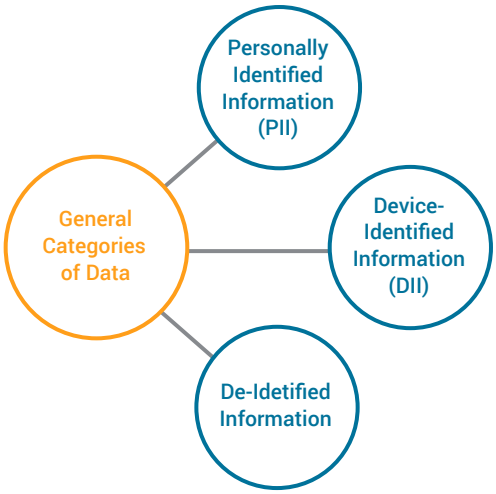
> Determine that the risk of re-identification of an individual is very small. In such cases, the risk of re-identification must be very small when the information is used alone, and must remain very small should the data be combined with other reasonably available information by an anticipated recipient to identify an individual who is a subject of the information.[2]

When evaluating re-identification risk under the Expert Determination rule, the statistician needs to consider the analytics approach, and corresponding privacy safeguards, used to de-identify and combine that data.

[1] HIPAA Journal, De-identification of Protected Health Information: How to Anonymize PHI, https://www.hipaajournal.com/de-identification-protected-health-information/
[2] HIPAA Journal, De-identification of Protected Health Information: How to Anonymize PHI, https://www.hipaajournal.com/de-identification-protected-health-information/

Network Advertising Initiative (NAI)

The NAI is a self-regulatory group that has developed standards for the use of data in digital advertising. More than 100 digital publishers and advertising technology platforms are members of the organization, and responsible participants in the digital advertising industry expect both members and non-members to comply with NAI guidelines. The NAI Code of Conduct is written specifically for digital marketing and considers digital use cases that could lead to the inappropriate use of an individual's sensitive health information.



The 2020 NAI Code of Conduct introduces a new concept, called Ad Delivery and Reporting or ADR. This specifically addresses the use of Sensitive Information, such as actual health data, in the analytics and measurement of digital advertising campaigns.

The NAI breaks out different types of data into three categories – Personally Identifiable information (PII), Device Identified Information (DII) and De-Identified Information. PII is defined as any data linked or intended to be linked to an identified individual, including name, address, telephone number, email address, financial account number, and non-publicly available government-issued identifier. Device-Identified Information (DII) is defined as 'data that is linked or intended to be linked to a particular browser or device.'[3] DII includes, but is not limited to, unique identifiers associated with users' computers or devices and IP addresses, even where such identifiers or IP addresses are not linked to PII.

The NAI defines De-Identified Information as 'data that is not linked or intended to be linked to an individual, browser or device.'[4] Unlike PII and DII, the use of De-Identified Information for ADR is permitted. Each category of data is mutually exclusive under the 2020 NAI Code of Conduct.

As we interpret the 2020 NAI Code of Conduct, health data cannot be connected to DII or PII without the opt-in consent of users, even when that data is anonymized.[5]

When evaluating the use of De-Identified Information, marketers should consider both how the data is de-identified and how that data could be used in analytics. If it is possible to take De-Identified Information and "walk backward" or use a crosswalk or other means to deduce the device ID or personally identified information, this should not be considered De-Identified Information under the NAI code. For De-Identified Information to remain pure and usable for ADR, the link between DII or PII and De-Identified Information needs to be broken.

[3] Network Advertising Initiative, 2020 NAI Code of Conduct, p. 19. https://www.networkadvertising.org/sites/default/files/nai_code2020.pdf
[4] Network Advertising Initiative, 2020 NAI Code of Conduct, p. 19. https://www.networkadvertising.org/sites/default/files/nai_code2020.pdf
[5] Network Advertising Initiative, 2020 NAI Code of Conduct, p. 19. https://www.networkadvertising.org/sites/default/files/nai_code2020.pdf

Re-identification Risks

A significant volume of research exists proving that seemingly anonymized data can easily be re-identified. Simply removing personal information from a set of data does not protect the privacy of individuals.

There have been numerous examples, not just in healthcare, of large anonymous datasets being re-identified. With the vast amount of personal data available online, and a huge marketplace for "scrubbed data" available for purchase, researchers have found it possible to combine publicly available information with "anonymized" data to re-identify individuals.

Consider, for example, researchers who re-identified Netflix users simply by analyzing their anonymized user ratings. In 2006, Netflix announced a software challenge, with a $1 million prize for the person or group who could improve the quality of its movie recommendation algorithm. To provide background data necessary to build the algorithm, Netflix published 100 million movie reviews posted by nearly 500,000 users between 1999 and 2005. The dataset was "anonymized" – personal identifiers like name and address were removed from the database.[6]

> Although the data contained no direct identifiers, within weeks of the data's release, two researchers were able to re-identify a subset of specific people by cross-referencing the Netflix data with IMDB.com ratings. Using just six ratings of obscure movies, the researchers re-identified individuals 84% of the time (if they were in both datasets). Including an approximate time the rating was made allowed identification 99% of the time.[7]

The researchers, Arvind Narayanan and Vitaly Shmatikov, concluded "…it is possible to learn sensitive non-public information about a person from his or her movie viewing history."[8]

---

[6] Narayanan, Arvind, and Shmatikov, Vitaly. Robust De-anonymization of Large Sparse Datasets. University of Texas at Austin. https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

[7] Lubarsky, Boris. Re-Identification of "Anonymized" Data. Georgetown Law Technology Review, 2017. https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/

[8] Narayanan, Arvind, and Shmatikov, Vitaly. Robust De-anonymization of Large Sparse Datasets. University of Texas at Austin. https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

While linking individuals to anonymous movie ratings is a violation of privacy, the sensitive nature of health information can make privacy breaches even more serious. The re-identification of Governor William Weld's personal medical information was the incident that spurred the de-identification framework in HIPAA.

In the mid-1990's, Massachusetts purchased health insurance for state employees and subsequently released records summarizing every state employee's hospital visits. Then-governor of Massachusetts William Weld assured the public that the data had been properly scrubbed. The fields containing explicit identifiers such as name, address, and Social Security numbers were removed, however, the record still contained almost a hundred unscrubbed attributes per patient. Latanya Sweeney, then a graduate student, obtained the data and used the Governor's zip code, birthday, and gender to identify his medical history, diagnosis, and prescriptions.

The most powerful tool for re-identifying scrubbed data is combining two datasets that contain the same individual(s) in both sets. Dr. Sweeney was able to re-identify Governor Weld's supposedly "anonymized" set of medical data by linking two databases together. She purchased the voter rolls from Cambridge, where Weld resided, then combined those rolls with the hospital data. Six people in Cambridge shared Weld's birthday, of those, half were men and only one lived in Weld's zip code. In this way she circumvented the scrubbing procedures and re-identified the "anonymized" data.

In a recent article in Nature Communications, researchers found that "even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization."

In fact, the study showed that by combining de-identified data with readily available sets of consumer data, researchers were able to re-identify the vast majority of patients.

99.98%
accuracy

Ability to re-identify an individual using 15 consumer characteristics.

79.4%
accuracy

Ability to re-identify an individual using four consumer characteristics.[10]

When two or more anonymized datasets are linked together, they can then be used to unlock other anonymized datasets. Once one piece of data is linked to a person's real identity, that data can then be used to destroy the anonymity of any virtual identity with which that data is associated. The ability to link even supposedly innocuous data exposes people to potential harm because of this.[9]

[9] Lubarsky, Boris. Re-Identification of "Anonymized" Data. Georgetown Law Technology Review, 2017. https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/

[10] Rocher, Luc, et al. Nature Communications, Estimating the success of re-identifications in incomplete datasets using generative models, https://www.nature.com/articles/s41467-019-10933-3

A Review of Analytics Approaches

1. Analytics by Agreement: Data Brokers

Traditionally, data brokers have invested in medical claims, prescription and electronic medical records that they de-identify and sell to pharmaceutical companies and others. Here the historical approach has been to secure de-identified healthcare data from Covered Entities (or Business Associates under HIPAA), aggregate the data into a central database, package, and sell the data. Under HIPAA, Covered Entities are defined as health plans, health care clearinghouses, and health care providers who electronically transmit any health information in connection with transactions for which HHS has adopted standards."[11]

As the importance of media and marketing measurement has grown, data brokers have retrofitted their analytics process to link more types of data together to provide a fuller picture of the patient.

Using a crosswalk file, data brokers match de-identified health records to de-identified media records. The different types of anonymized data are brought together and combined in a central database. But, because the data is anonymized before it is matched, the ability to accurately match records decreases. While this reduces the risk of re-identification, it also lessens the utility of the data.

When data is matched using a match key (or crosswalk) that links device ID or digital IDs to health tokens, technically the data could be walked back and reconnected. This increases the chance that medical records could be linked back to an individual. Data brokers attempt to reduce that risk by keeping health data separate from media data, including storing at different companies and then agreeing, through contractual commitments, to not re-identify. However, because the end analytics product is not provided at the person-level, the data broker – but not the client – can technically connect an individual to personal health information. The ability to do so means the process is not compliant with the 2020 NAI Code of Conduct for ADR because the Device Identified Information, which is linked to an individual, could be tied back to personal health information. This introduces significant risks of re-identification into the ecosystem because of the potential that a nefarious or incompetent actor could expose personal health information.

2. All Under One Roof: Technology that Enables De-Identification and Data Combinations for In-House Analytics

Certain new technology companies seek to connect and sell de-identified health information either by providing technology to third parties to make data linkages or by providing a marketplace of available datasets. These companies are now making available media and consumer data as well, raising serious privacy concerns.

Similar to legacy data brokers, these approaches result in scenarios with enough data accessible within the same company to potentially allow for re-identification. As described above, having de-identified health data combined with digital campaign metadata still creates a unique fingerprint for a given record. (**Figure 1**) When that same metadata exists within the same company while connected to device IDs, it is incredibly easy to re-identify. Similar to the approach above, given the ability to tie health data to DII, the approach should not be considered compliant with 2020 NAI Code of Conduct.

[11] HIPAA Privacy Rule, Information for Researchers. US Department of Health and Human Services, National Institutes of Health. https://privacyruleandresearch.nih.gov

Figure 1. In this example, a data aggregator sends their client (a pharma brand) media exposure data with health data appended.

HEALTH PROXY ID	DATE	BRAND	PLACEMENT	SITE	AD CREATIVE
hv235wd	10/1/2021				
hv235wd	10/10/2021	Brand A	Brand A_200X250	Med Site B	Creative A1
hv235wd	10/25/2021				

In **Figure 1**, the combination of Date, Brand, Placement, Site, and Ad Creative with the assumption that each person sees an ad two times results in 19.2 trillion unique combinations:

(Date (365 values) x Brand (10) x Placement (20) x Site (20) x Ad Creative (3)) ^ Impressions (2) = (4,380,000)² unique combinations.

This means that more than 99% of the individuals represented in the data set can be uniquely identified by their metadata. If combined with the ability – directly or indirectly – to look up metadata combined with a digital identifier such as a device ID, it is clearly possible to then link the device to health data. An example of this could be the use of a data management platform (DMP) which stores digital IDs along with campaign metadata.

3. Distributed Analytics - Leveraging Technology to Protect Privacy

With a distributed approach to analytics, there is no need to combine all necessary data in a single location. Analytical techniques are applied on-site at the source of the data, thereby maintaining individual privacy while still extracting analytical insights. Distributed approaches, also sometimes known as federated approaches, are increasingly being leveraged as a means to protect privacy while still extracting value from analytics (**Figure 2**). For example, Google is using this approach to create predictive models for applications, including Google maps and predictive search. They recently depicted this approach here: https://federated.withgoogle.com/.

In the case of analytics of digital healthcare campaigns, this approach is used to break the link between actual health data and any other identifiers or metadata that could be used to re-identify an individual. In doing so, it becomes impossible to "walk backward" and connect a health record to an individual or device.

Features of this approach include:

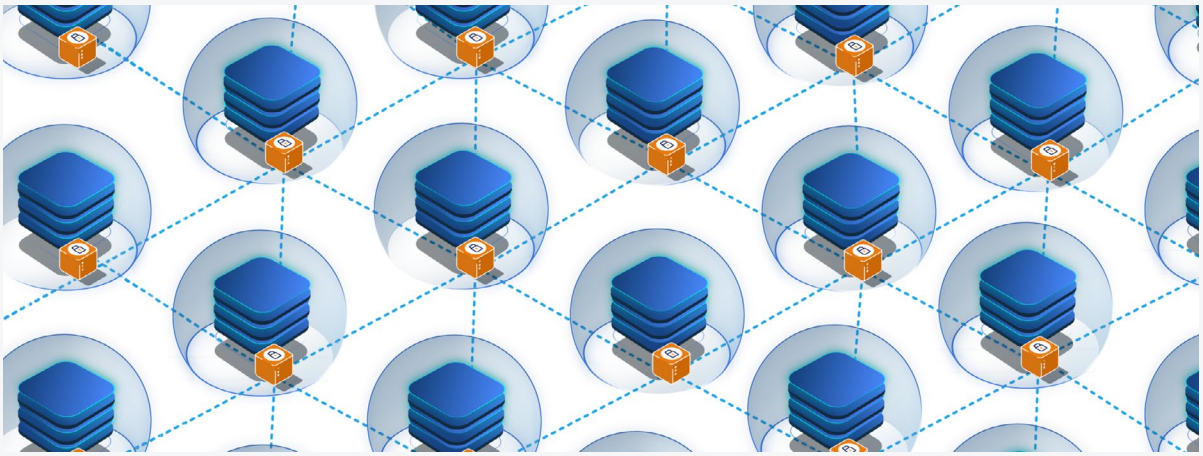
- Digital identifiers are never stored in the same location as health data
- Campaign metadata (described above to be uniquely identifying) is never connected to health data
- Analytics are only provided on large groups of people

Veeva Crossix uses a distributed approach with our Crossix SafeMine technology. SafeMine is deployed behind the firewalls of HIPAA covered entities, such as pharmacies, health plans, PBMs, EMR companies and clearinghouses. For digital analytics, only lists of hashed IDs representing people who saw an advertisement are sent to these installations. Once the hashed IDs are linked to health data, the digital ID is removed and replaced with a token, a de-identified ID that cannot be linked back to any digital information.

In this process, no metadata is ever combined with health data. Metadata is used only to create large cohorts of people. For example, metadata would be used to create a group of people who saw an ad on a given website.

By breaking the link between digital data (digital identities and metadata) and health data, the distributed approach represents the gold standard for privacy-safe analysis of digital health campaigns. A high velocity, high accuracy, technology-enabled modern approach to privacy-safe analytics, nearly eliminating risks of re-identification in the process.

Figure 2. Analytical techniques are applied on-site at the source of the data, behind privacy firewalls. Personal information never leaves the secure environment of the covered entity.



Conclusion

In today's privacy-focused environment, protecting confidential patient information should never be dependent on contractual obligations alone. Instead, the technology used to process secure consumer and patient data must have inherent privacy protections that go beyond HIPAA requirements. Responsible players in the ecosystem must understand and adhere to these best practices to allow for the extraction of value toward better business and health outcomes, while protecting the ecosystem by protecting the privacy of patients.

About Veeva Systems
Veeva is the global leader in cloud software for the life sciences industry. Committed to innovation, product excellence, and customer success, Veeva serves more than 1,100 customers, ranging from the world's largest pharmaceutical companies to emerging biotechs. As a Public Benefit Corporation, Veeva is committed to balancing the interests of all stakeholders, including customers, employees, shareholders, and the industries it serves.
For more information, visit **veeva.com**.

Veeva Systems
Global Headquarters
Pleasanton, California, USA
4280 Hacienda Drive
Pleasanton, California 94588
+1 925 452 6500 | veeva.com/contact | veeva.com

Learn more at veeva.com | 925-452-6500 | veeva.com/contact