Health Care Vendor Data Management in the Era of Big Data and Machine Learning

A Closer Look at the Risks Covered Entities Must Manage when Outsourcing Data Analyses to Vendors

Daniel Fabbri

The digitization of patient records and push for interoperability has enabled unprecedented data analyses of health care data, which have the potential to improve patient care and treatment. Covered entities are turning to vendors with expertise in areas like billing, population health, readmission modeling, etc. to analyze and leverage their data sets. Often, using a vendor means sending data to the cloud.

Third-party vendors are increasingly moving to cloud-hosted solutions. A 2017 HIMSS Survey showed almost 89 percent of health care respondents were using some form of vendor-hosted solution,[1] and a 2018 BCC Research report predicts health care cloud computing will grow at a compounded annual growth rate of 11.6 percent through 2022.[2] These vendors ingest data, after agreeing to a business associates agreement (BAA), and perform their intended analyses using the data.

While there are many benefits from these vendors, there are additional risks that covered entities must understand when outsourcing data. The Health Insurance Portability and Accountability Act (HIPAA) makes the covered entity responsible for "the use and disclosure of individuals' health information." A covered entity is not only tasked with monitoring how patients' data are accessed within their digital network but also protecting any data sent outside their network controls.

This article addresses the risks that covered entities must manage when outsourcing data analyses to vendors, with a focus on how these risks change in the era of vendor-hosted solutions, big data, and machine learning. Specifically, this article discusses the risks of data mixing, machine learning model mixing, and data repurposing. Finally, the article addresses improvements that are needed to modernize current external

Daniel Fabbri, PhD, is the founder and chief executive officer of Maize Analytics and Assistant Professor of Biomedical Informatics and Computer Science at Vanderbilt University.

data governance practices to ensure that sufficient monitoring processes are in place.

Sending Data to the Cloud

The migration to cloud-hosted solutions makes sense for many covered entities. Covered entities no longer must manage their hardware resources (and their failures), can scale computational resources on-demand, and can distribute data geographically for data recovery. However, the way that covered entities and their vendors leverage the cloud can vary.

Some covered entities are beginning to spin up their own clouds and allow vendors to deploy solutions within it. Thus, vendors get the benefits of cloud infrastructure, while covered entities have better visibility into where data go, which ports are open on a server, and can manage access controls. Moreover, these covered entity-managed clouds ensure a logical separation between their data and another health care organization's data.

In contrast, vendor-managed clouds are often a black box for covered entities. Most of the time, once data are transferred into the cloud environment, it is out of sight of the covered entity. The covered entity loses the ability to track how the data are used, where the data are stored, and what other data are combined with the covered entity's data. The covered entity is at the mercy of the vendor's security controls and data management processes. Only contractual agreements bind what vendors can do with covered entities' data, and those controls are often hard to check.

Given this risk, why do covered entities increasingly leverage vendor solutions in vendor-managed clouds? First, new big data processing systems require expert skills to manage, which may not be available in each covered entity's organization. Second, these solutions increasingly encompass complex data pipelines that are difficult to manage without full control of the environment, thus limiting vendors' appetite to deploy in covered-entity-managed clouds. Third, the economies of scale decrease as vendors manage their solution in multiple locations, the cost savings of which can be passed on to the covered entity.

In part due to this growing reliance on vendors, The National Institute of Standards and Technology (NIST) released revision 2 of NIST SP 800-37, Risk Management Framework (RMF) for Information Systems and Organizations in December 2018. A major objective of the update was to integrate security-related, supply chain risk management concepts into the RMF.[3] The revised RMF focuses organizations on the need to manage risk in their supply chains,[4] which are often associated with an "organization's decreased visibility into, and understanding of, how the technology that they acquire is developed, integrated, and deployed."[5]

Given this push to vendor-managed clouds, covered entities must understand their risks and approaches they can take to protect their data.

Data Mixing

After data are sent to a vendor-hosted solution, the vendor controls where the data are stored. Data can be stored in a number of different ways, including flat files (*e.g.*, CSV), relational databases, key-value stores, distributed file systems (*e.g.*, Hadoop), and bucket stores (*e.g.*, AWS S3). Data are often transformed from one format to another as data are analyzed and processed for value (*e.g.*, from an AWS S3 bucket into a relational database). A key design question that covered entities should ask no matter the storage system that is used is: what is the system's tenancy?

A single-tenant data storage architecture ensures a logical separation between each covered entity's data. On Amazon Web Services or other cloud providers, single tenancy can be guaranteed with virtual private networks (VPNs) and organization-specific data stores. As a result,

data from two organizations can never be mixed on a file storage system or within a database.

For additional separation, dedicated hardware ensures that two virtual applications never run on the same machine. Dedicated hardware protects against reported vulnerabilities like Meltdown and Spectre, in which keys and passwords were inappropriately read across virtual machines running on the same hardware through a shared processor and memory state.

In contrast, a multi-tenant architecture stores two or more covered entities' data in the same data store. For example, data from organization A may be stored in the same database as organization B. Vendors use various software access controls to ensure each organization only sees their data. Vendors can prefer multi-tenant architectures because they only need to support a single environment instead of one per customer.

The risks of multi-tenancy can be large. Simple software bugs can result in one covered entity seeing another's data. For example, if a new developer forgets to limit the patients shown on a Web page to a single organization, users can see every patient, even if they are not from their organization. Innocuous bugs can have profound implications.

In June 2017, the Office for Civil Rights (OCR) issued a newsletter outlining what covered entities and third parties should consider when implementing cloud computing platforms.[6] In this newsletter, the OCR states, "misconfigurations of file sharing and collaboration tools, as well as cloud computing services, are common issues that can result in the disclosure of sensitive data, including ePHI."[7] A covered entity's risk management and risk analysis process should be aware of the risk of these misconfigurations in multi-tenant environments.

Steps to Take: Ask if data will be in a single or multi-tenant environment. If multi, ask what controls are in place to prevent inappropriate data mixing.

Machine Learning Model Mixing

The health care community has shown increasing interest in machine learning and artificial intelligence over the last five years. Recent work in industry (*e.g.*, Google's work for retinal disease[8]) and academia (*e.g.*, Vanderbilt's work on readmission prediction[9]) have demonstrated modern algorithms' abilities to learn from historical examples and train a model to predict readmission or determine a diagnosis. To improve model effectiveness, covered entities and vendors continuously look to build larger and larger training data sets.

One way to get more data for machine learning is by mixing data from multiple covered entities. As described above, multi-tenancy easily allows a vendor to aggregate multiple entities' datasets for machine learning. When multi-tenancy is restricted, vendors can deploy a single-tenant application, but copy the data into a data warehouse, effectively aggregating every entities' data for machine learning training. Because covered entities have no visibility into the vendors' data store, such a copy-and-train approach is unfortunately easy to execute.

When mixing data is not an option, an alternative machine learning approach is to iteratively train the model on multiple data sets. In an iterative training approach, data from organization A are first used to train a model, then that model is transferred to organization B's environment, and then B's data are used to train the model further, and then C, and so on (*e.g.*, see algorithms like stochastic gradient descent for iterative training). Thus, while data are not mixed between organizations, the resulting model that is produced relies on data from multiple organizations. The resulting model is then used in each organization for prediction and classification (*e.g.*, predicting readmission or disease).

There are multiple risks that covered entities must understand when using machine learning models that are trained on multiple entities' data.

First, the underlying data distributions between each covered entity may be different, therefore resulting in incorrect predictions. For example, if organization A has a higher diabetes rate than organization B due to its ethnic patient makeup, a model trained on data from A (or A and B) may over-report diabetes.

Second, sharing the model has the risk of exposing patient data from one covered entity to another. For example, consider if organization A has a celebrity patient that is white, age 50, and has a history of diabetes, an aorta heart valve replacement, and AIDS. A curious user at organization B could use the model to determine if the celebrity has AIDS by entering the other characteristics (if "white, 50, diabetes, aorta heart value" implies "AIDS"). While data obfuscation methods like k-anonymity, l-divergence, and differential privacy can help mitigate inadvertent exposures, anonymization methods are not always used or deployed correctly.

Third, shared models have the risk of making incorrect conclusions because of different semantics between covered entities. For example, if Oncologists at organization A are in the Oncology Department but in the Oncology Service in organization B, then models that look at medications from the Oncology Department would be incorrect for B. These mismatched semantics exist everywhere across organizations from department names to diagnosis codes to clinical note text. Without a robust mapping system, some models may produce incorrect or nonsensical results.

Fourth, shared machine learning models may not consider the differing policies that are in place at each covered entity. Consider organization A that allows its employees to access their own medical record and organization B does not. If trained on data from both, will a privacy monitoring system learn to mark all self-access as inappropriate or appropriate? In this shared model, it is unclear if a covered entity's self-access policy will be correctly followed.

Fifth, as new covered entities are added to the vendor's customer base and used to iteratively train the model, the model's predictions can change. As a result, any previous validation work may no longer be valid and may need to be re-done. Continuous validation processes are necessary to ensure continuous prediction quality as the system is updated.

Steps to Take: Covered entities should ask what data are used to train the model and if the model is iteratively updated over time or static. If the model is trained on data from multiple entities, ask how and if data are mixed.

Data Repurposing

Covered entities send data to vendors to perform a service but often can only rely on contractual obligations to ensure the vendor only uses that data for contracted purposes. However, given the lack of visibility most covered entities have into their vendors' data storage system, it is, unfortunately, possible for vendors to repurpose the data set and leverage it to build new products and services. If a vendor utilizes ePHI outside the scope of the business associate agreement, the covered entity could be subject to liability for a HIPAA violation.[10]

To reduce the chances of data repurposing, covered entities should work to improve their visibility into their vendors' data processing system to ensure ePHI is being used within the scope of the BAA. One way to improve visibility is to require vendors to send the covered entity all query logs (*e.g.*, from relational databases) from their environment. By monitoring database activity and looking for ad-hoc queries instead of automated jobs, organizations can start to see if unexpected operations are being performed.

Additionally, covered entities should request an accounting of every location their data may be stored in the vendor's cloud. The provenance from file to relational database to web application is essential to track and determine the risk to patient data.

Steps to Take: Covered entities should update contracts to allow them to request activity logs and data storage architectures from vendors. This information allows covered entities to better monitor how vendors use covered entities' data.

Data Governance

Covered entities must adhere to the principle of data minimization and limit the ePHI distributed to vendors to the least amount necessary for them to perform a function.[11] Performing good data governance means continually evaluating risk, tracking where data are sent, and the type of information that is disclosed.

While the data governance process should be an ever-evolving practice of understanding risks, at many institutions, this practice is often reactionary, where data governance processes are only evaluated or updated after misuse or a security breach is identified. Furthermore, the data governance approval process often is a one-time approval process. This means data feeds to each vendor get approved during implementation but are not typically monitored in an ongoing way.

To develop a robust external data governance program that manages multiple vendor-hosted solutions, covered entities should have an accounting of each data feed, including the fields that are sent (*e.g.*, MRN, SSN, DOB, etc.), the list of patients transmitted, and field additions over time. Without these details, if a vendor were to be breached, covered entities fall back to remediating their entire population instead of the smaller subset of patients the vendor received.

Steps to Take: Covered entities should develop processes to continuously review and monitor ePHI files being sent to vendors. Additionally, covered entities should be able to identify which patient's ePHI is sent to each vendor.

Guidance

How can covered entities mitigate risks while utilizing vendors for their data analyses? When looking at a potential vendor, covered entities should conduct a risk assessment that goes beyond information security controls and addresses the issues listed above. The "Steps to Take" outlined above are an initial list of questions covered entities should ask a potential vendor. From there, a BAA needs to address issues identified in the risk assessment and ensure controls, both contractual and technical, are in place. Throughout the covered entity and vendor relationship, a covered entity should have a process in place for continuous auditing of data sent out, and the risks associated to the data.

As cloud-hosted solutions and machine learning gain popularity, it is important for covered entities to understand the associated risks and how best to mitigate them. While most vendors can be trusted to do the right thing, covered entities should have the necessary legal and technical processes in place to identify issues.

Endnotes

1. HIMSS Analytics 2017 Essentials Brief: Cloud (2017) *www.himssanalytics.org/sites/himssanalytics/files/Cloud%20Study_2017%20Snapshot.pdf*.
2. New BCC Research Report Estimates Global Market For Healthcare Cloud Computing to Reach $35.0B by 2022 (January 2018) *www.bccresearch.com/pressroom/hlc/new-bcc-research-report-estimates-global-market-for-healthcare-cloud-computing-to-reach-$350b-by-2022*.
3. Nat'l Inst. of Standards & Tech., U.S. Dep't. of Commerce, NIST Special Pub. 800-37, Rev. 2, Risk Management Framework for Information Systems and Organizations vi (Dec. 2018) [hereinafter NIST SP 800-37], doi.org/10.6028/NIST.SP.800-37r2.
4. Nat'l Inst. of Standards & Tech., U.S. Dep't. of Commerce, NIST Special Pub. 800-37, Rev. 2, Risk Management Framework for Information Systems and Organizations 20 (Dec. 2018) [hereinafter NIST SP 800-37], doi.org/10.6028/NIST.SP.800-37r2.

5. *Id.*
6. U. S. Department of Health & Human Services, Office for Civil Rights, File Sharing and Cloud Computing: What to Consider (June 2017), *www.hhs.gov/sites/default/files/june-2017-ocr-cyber-newsletter.pdf*.
7. U. S. Department of Health & Human Services, Office for Civil Rights, File Sharing and Cloud Computing: What to Consider (June 2017), *www.hhs.gov/sites/default/files/june-2017-ocr-cyber-newsletter.pdf*.
8. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs (December 2016) *jamanetwork.com/journals/jama/fullarticle/2588763*.
9. Predicting Negative Events: Using Post-discharge Data to Detect High-Risk Patients (February 2017) *www.ncbi.nlm.nih.gov/pmc/articles/PMC5333334/*.
10. 45 C.F.R. § 164.502(a)(3); 45 C.F.R. § 160.402(c).
11. 45 C.F.R. § 164.502(b).

Reprinted from Journal of Health Care Compliance, Volume 21, Number 2, March–April 2019, pages 19–24, with permission from CCH and Wolters Kluwer.
For permission to reprint, e-mail permissions@cch.com.