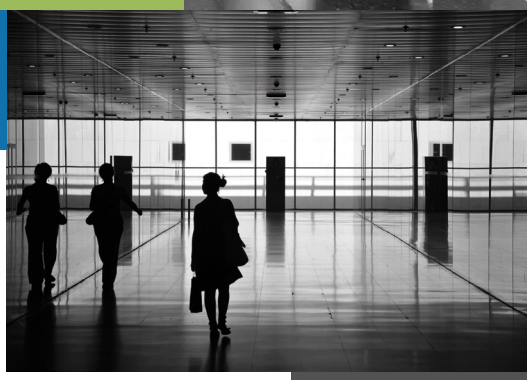
PRIVACY ANALYTICS

WHITE PAPER

Unlocking Big Data for Healthcare

PHI is the most sensitive of all data – and unlocking it for Big Data Analytics is no easy feat. The risk of re-identification grows along with the size of health data being created. Fears around patient privacy hinder progress. This white paper discusses the inherent risks in large and linked datasets and offers guidance on the steps companies can take to protect patients and themselves.



PRIVACY ANALYTICS

a QuintilesIMS company

Executive Summary

There are powerful opportunities facing healthcare today. Big Data Analytics (BDA) is being leveraged by a broad spectrum of healthcare organizations to solve ongoing challenges. BDA is opening doors when it comes to minimizing expenses, saving time, improving patient health and advancing precision medicine. With such obvious benefits, it should come as a surprise that BDA is still in its infancy.

PHI is the most sensitive of all data – and unlocking it for BDA is no easy feat. Fears around patient privacy hamper progress. As health data repositories grow, more and more patient data is being linked from multiple sources. The risk of re-identification grows along with the size of health data being created. Petabytes of health data are sitting in siloes that could hold important answers to improving patient outcomes and minimizing the cost of healthcare delivery. Storage of the data is the easy part. Deciding how to unlock the data so it can be used for analysis or shared with an ecosystem of partners is the greater challenge. De-identification allows for the unlocking of health data for secondary use, but only a risk-based method can address the sensitivity and multitude of data. Understanding the privacy risks and finding the balance between data access and data anonymity is central to BDA's future success.

Health Information Systems (HIS) vendors are in a unique position to capitalize on their data assets and deliver on the promise of BDA. With data on billions of healthcare claims and millions of patients, they have the foundation to build the integrated data repositories needed to generate new insights into healthcare's biggest challenges. This white paper looks at how BDA can influence healthcare for the better, discusses the inherent risks in large and linked datasets, and offers guidance on the steps companies can take to protect patients and themselves.

The Future of Healthcare Relies on Big Data

Healthcare is undergoing a seismic transformation. Mounting financial pressures and changing clinical practices are transforming the way that we approach healthcare and manage patient outcomes. Together, these forces are driving greater demand for Big Data Analytics (BDA).

The benefits of BDA are not hard to imagine. BDA allows healthcare stakeholders to gain insights on how to avoid unnecessary costs of care while simultaneously improving patient outcomes. It relies on massive real world datasets that contain information on millions of patients to reveal the patterns, correlations and trends hidden within the data. BDA provides the potential to:

- **Minimize costs overall.** Using trends and predictive analytics, hospitals can find wasted expenditures and allocate financial resources where they are really needed. The Institute of Medicine reported in 2012 that $750 billion is wasted each year on unnecessary care that fails to make anyone healthier[1].
- **Detect disease at earlier stages in their progression.** By catching diseases early on, it is possible to prevent hospitalizations and complications, two significant factors that increase the costs of care.
- **Improve population health.** Chronic conditions are becoming more prevalent. By gaining a better understanding of the etiology of diseases like diabetes and cancer, it is possible to identify populations that are at risk. This can lead to earlier interventions and inform policies to support healthier lifestyles.
- **Advance precision medicine.** By applying predictors found in patient populations it is possible to identify individuals who may be susceptible to behaviors that negatively impact their health, such as medication non-adherence. BDA can also identify which interventions will most likely lead to a successful outcome.

Unlocking the value of BDA is no simple task. The sheer amount of data being stored is growing exponentially. According to one research firm, 30% of the world's data is health data[2]. Healthcare data is also the fastest growing. Each year it grows by 48%[3]. The move from paper-based systems to electronic health records (EHRs) has created an explosion in the volume of patient data being generated. This not only creates issues around the management of patient data, it also means that the data exists in many different formats from different EHR systems. Dealing with the growing volume and variety of data demands scalable solutions that are designed to manage Big Data.

The digitization of EHRs also means we are seeing more data linked. Health data is being linked to social media and wearables data. Financial data is also tied to health records via insurance claims and payments. All data is quickly becoming PHI. By using this data for

secondary purposes, privacy concerns rise. BDA could be undermined if concerns about privacy are not addressed. Patient privacy has been a major topic of discussion in conversations on BDA in healthcare[4]. It has even been said that healthcare privacy could crash Big Data if it's not done right[5]. Knowing how to manage risks from sharing Big Data for secondary uses and how de-identification — when done properly — can protect the privacy of patients will help to further BDA's use in healthcare. The benefits of BDA are considerable, but until healthcare organizations can surmount the barriers, they will stay out of reach.

The Shift from Volume-Based to Value-Based Care

Healthcare is facing a looming fiscal crisis. Recent news from the Medicare Trustees' report is that the fund will be insolvent by 2028[6], right around the time the last of the baby boom generation is set to retire. This is being driven, in part, by the fact that seniors are more prone to costly chronic health conditions, like diabetes, arthritis and cancer. Chronic diseases are among the most common, costly and preventable of all health problems.

It is forcing governments to shift from traditional volume-based care models to new models of value-based care. Until now, healthcare payments have followed a fee-for-service mechanism, where providers are paid based on the number of tests ordered, patients seen and procedures completed, regardless of patient outcomes. With national healthcare expenditures totalling more than $3 trillion dollars[7], there is an urgent need to gain control over costs. The old 20th-century model of medicine based on volume, the 'fee-for-service,' is no longer viable. Value-based care is about delivering better health outcomes at a lower cost, not simply delivering more care.

Using BDA to Improve Medication Adherence Could Save Billions

Self-management of health conditions is a challenge, particularly for elderly patients who may have multiple chronic conditions and take dozens of pills a day to manage them. Poor adherence to prescribed medications is common. Studies have shown that as many as 50% of patients fail to take their medications as prescribed[8] even though the lack of adherence is associated with poor therapeutic outcomes and disease progression. The result is higher healthcare costs due to more visits to the doctor and increased use of urgent care facilities. The economic impact of non-adherence is estimated to be between $100 billion and $289 billion annually[9]. Improving medication adherence rates would reduce hospital admissions and lead to cost savings.

This is one area where BDA can provide value. Improving medication adherence can create significant savings and result in better patient health — objectives consistent with value-based care. While there are many ways to look at the problem of non-adherence, all of them leverage data analytics principles. This includes segmentation and comparison of non-adherent and adherent populations, evaluation of alternative therapies, analysis of test results examining different approaches to improve adherence, and applying these models to identify patients at risk and choose the most

appropriate intervention plan for them. By analyzing the behaviors, interactions and outcomes of patient groups, it is possible to identify the therapies and interventions most likely to result in optimal health for an individual.

Performing this type of analysis is contingent on the availability of "Big Data" — extremely large datasets that contain extensive patient information. Ironically, it is not the organizations involved in the delivery of primary care or those doing drug development that are in the best position to deliver on Big Data.

Many major health information systems (HIS) vendors are part of the labyrinthine network of data applications that deliver healthcare's many functional requirements. These are the organizations that provide clinical software, like EHRs and computerized physician order entry (CPOE) systems, as well as administrative software for medical coding, billing, and claims reconciliation are a part of the complex reimbursement process. These HIS vendors have patient datasets that are characterized by the three "V's" of Big Data — volume, velocity and variety — that makes them ideally suited for BDA.

While this puts HIS companies in a unique position to be able to capitalize on the opportunities for BDA, sharing data with internal research groups or external organizations requires companies to mitigate the chances that a patient could be re-identified from their data. The next section looks at the privacy issues raised when data sources are linked.

Data Linking Adds Variety and Value for BDA

Finding patterns and trends in data that are widely applicable requires a dataset that reaches across state lines and draws from multiple touch points along the continuum of care. It is why demand will be greatest for those data repositories that make the most of linking together information from multiple sources. This includes:

1. Aggregating data from healthcare providers around the country so that a greater portion of the population is represented; and,
2. Connecting data from EHRs and disease registries with other administrative, pharmaceutical and claims data to provide a greater amount of detail on a patient's care experience.

It is the second aspect, creating a comprehensive view of the patient's journey, which creates concerns around patient confidentiality. Linking data expands opportunities for data's use in BDA but also creates privacy concerns. The more information available about an individual, the more likely it is that identifiers within the data could be used to re-identify that person. Identifying pieces of information like zip code, gender and age can be used in combination to zero in on a unique individual. Once identified it is possible to learn even more about the individual through the other data linked to their record.

To be useful for Big Data, however, databases need to offer not only very large volumes of records but also provide access to a wide variety of data. The industry is experiencing greater

demand for data offerings that are able to bridge the gaps between different information domains and tie together clinical, research, administrative and social data. To enhance the value of their datasets, HIS vendors can use their clinical and claims data as a foundation for linking to other data types that are useful in BDA, such as:

- Clinical research and medical reference materials
- Clinical data contained in unstructured formats, such as physician's notes
- Assessment Results in semi-structured formats
- Health product (e.g. drug information) data
- Genomics and gene sequencing data
- Data from wearables (Fitbits) and other wireless monitoring devices
- Social networking data (Twitter, Facebook) and search engine data (Google)

The linking of data is becoming more important for secondary uses. Incorporating diverse sources of data that capture information from different points along the cycle of care adds tremendous value for BDA. Linked datasets extend the breadth of information provided on each patient and can offer insights into a patient's behavior and lifestyle choices, the progress of disease from diagnosis through to outcomes, and their physical, mental and emotional state along the way.

The responsible sharing of Big Data begins by taking steps to ensure that protected health information (PHI) that could be used to identify a patient is removed or generalized. Data de-identification is a critical step to protect both patients and data owners. Failing to take appropriate precautions to prevent a privacy breach can result in legal and financial consequences as well as do irreparable harm to a company's reputation.

While there are various approaches to anonymizing patient data, the Expert Determination, or Statistical Method is standing out as the optimum approach. Many regulators and standards are emerging to reflect a shift to a risk-based approach to de-identification, including the Institute of Medicine, HITRUST Alliance, PhUSE and European Medicines Agency. This risk-based method of de-identification effectively protects privacy because it can address numerous quasi-identifiers in the data without wiping out its utility for analysis. Expert Determination renders data anonymous so that PHI can be used safely and also retains the data's inherent quality. The availability of software to fully automate de-identification using this risk-based methodology also makes it a reasonable solution for data on a massive scale because it can address data volume and variety.

Using Expert Determination for Big Data and Privacy

Big Data creates rich and detailed sources of information for analytics but this comes with the requirement to manage and protect the data. Healthcare datasets are replete with sensitive information about patients. If a dataset contains information about a stigmatizing health condition or other highly sensitive information, it could have devastating consequences for a patient. In the

past, patients — such as those living with HIV or AIDS — have suffered discrimination as a result of their health information being disclosed[10].

In the U.S., the HIPAA Privacy Rule governs the standards for the use and disclosure of PHI held by covered entities (health plans, providers and healthcare clearinghouses) and their business associates (those who use or provide PHI in providing services to covered entities). Under HIPAA's Privacy Rule, data can be de-identified to remove PHI using either the Safe Harbor or Expert Determination standard. With PHI removed data is no longer subject to HIPAA restrictions since it is considered anonymous[11].

The challenge with linked data is that the aggregation of multiple datasets means there is a significant amount of PHI in the data. It will include many direct identifiers like Social Security Number, health insurance number or username. It will also include numerous indirect identifiers like postal code, profession, date of birth, date of admission, and diagnosis codes. De-identifying data requires masking or removing direct identifiers, to eliminate the possibility of these fields being used to easily re-identify someone. Indirect identifiers, on the other hand, are useful for analysis and need to be retained. This is the data that can provide valuable insights into regional variations, socioeconomic impacts or behavioral influences on health. While these fields will be de-identified, it is desirable to keep as much specificity as possible in these data elements for analytic purposes. However, the more indirect identifiers there are associated with an individual, the easier it will be to re-identify that person. In this situation, de-identifying data to a high standard cannot be achieved using the Safe Harbor method which is focused on removing 18 specified data elements. Responsible data sharing requires the use of a risk-based approach, like Expert Determination, which relies on expert use of statistical principles to render information non-identifying. Expert Determination scientifically measures the re-identification risk in the data from the presence of indirect identifiers. De-identification techniques (generalization, aggregation, shuffling and randomization) can then be applied to reduce the data's identifiability to a degree consistent with precedents set out by reputable data organizations, like the Centers for Disease Control. The ability to reliably anonymize data while retaining high data quality is the reason that leading data organizations from around the world, like the Institute of Medicine, HITRUST, PhUSE and the Canadian Council of Academies, have all recommended the use of a risk-based approach, like Expert Determination, to de-identify data.

Determining how to address privacy issues is one of the major barriers to organizations moving ahead with BDA initiatives. Companies that want to pursue opportunities with BDA need to establish strong privacy practices in addition to implementing risk-based data de-identification. The following section outlines some best practices for creating a privacy-focused culture in your organization.

Best Practices to Create A Privacy-Focused Culture

A survey from the Office of the National Coordinator for Health IT in 2014 showed that patients are growing more comfortable with the idea of sharing their data for secondary uses but trust is still a major issue[12]. Companies that are

serious about BDA will need to implement best practices around privacy and security that not only protect patients but also protect their reputation as a responsible data owner. The following strategies can help data providers take a proactive approach to privacy risk that allows them to confidently move ahead with BDA initiatives. This includes:

1. Creating a Privacy Task Force

This is a cross-functional group that includes members from multiple disciplines in the organization. It is generally made up of the Privacy Officer, IT staff and business representatives, among others. The objective of the group is to assess internal and external privacy threats and establish risk mitigation strategies. They meet regularly to ensure the Chief Risk Officer and Chief Legal Officer are apprised of plans for using or sharing data for BDA.

2. Establishing frameworks to assess risk exposure

Assessing who will have access to the data, how it will be accessed and used, and where it will be stored will be on a case-by-case basis. To quantify your organization's risk exposure in each situation, you need to look at both the data and the data recipient. A framework allows you to objectively assess the context in which data will be shared. It looks at factors such as the motives to re-identify the data, the security controls in place for the data recipient and the sensitivity of the data. Scoring mechanisms quantify the level of risk. Guidance is then given on how extensively de-identification should be applied to sufficiently anonymize data for the given context.

3. Building employee expertise

Employees who will have responsibility for dealing with data requests and/or authority to release data should receive formal training about privacy issues and de-identification. It allows key staff to become familiar with the standards and techniques of de-identification and gain skills in managing re-identification risks.

4. Leverage software for scalability

The amount of data and number of fields in these datasets are making it impossible to manage de-identification with rules-based approaches or masking tools. The tasks of measuring risk and performing de-identification can be handled with the use of automated de-identification software. Software solutions are now available that employ the expert statistical methods needed by HIPAA's Expert Determination standard. They can scale for the amount of data and data variety, eliminating two major hurdles to performing quality BDA.

Implementing strategies like those listed above will move your company towards organizational readiness for BDA. Putting procedures and safeguards in place not only demonstrates the desire to comply with privacy regulations, it also establishes a defensible process in the event of a privacy audit or lawsuit resulting from a data breach.

CONTACT US

251 Laurier Ave W
Suite 200
Ottawa, Ontario, Canada
K1P 5J6

Phone: 613.369.4313

www.privacy-analytics.com

sales@privacy-analytics.com

Conclusion

We are just starting to realize the promise that BDA holds for the future of healthcare. While the benefits are great, there are barriers to unlocking the potential. These challenges are not unsurmountable, however, and as this market matures, opportunities to aggregate and share data will proliferate. With more organizations aiming to take advantage of powerful analytic tools and techniques available, we will see healthcare organizations not only addressing escalating healthcare costs and improve healthcare outcomes for patients - but overcoming them.

Leveraging this data has real value; the McKinsey Global Institute has estimated that Big Data could be worth $300 billion to the US healthcare industry from improved quality and efficiencies[13]. Organizations need to start now to establish strong privacy practices that will allow them to take maximum advantage of their wealth of data. Following the best practices of privacy-focused organizations and taking steps to establish de-identification practices based on the Expert Determination standard will enable HIS vendors to turn their information into a strategic asset for BDA.

To learn more about the application of Big Data Analytics, make sure to download our case study on ASCO CancerLinQ. With over one million patient lives, the CancerLinQ portal is the world's only learning health system for oncology research.

Sources

1. Kliff, Sarah. (2012, September 7). We spend $750 on unnecessary health care. Two charts explain why. The Washington Post. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2012/09/07/we-spend-750-billion-on-unnecessary-health-care-two-charts-explain-why/

2. Ponemon Institute. (June 2008). Survey on the Government of Unstructured Data. Retrieved from: https://varonis-assets.s3.amazonaws.com/pdfs/Unstructured_Data_Governance_Study_by_Ponemon.pdf

3. EMC Digital Universe with Research and Analysis from IDC. (2014). The Digital Universe Driving Data Growth in Healthcare: Challenges & Opportunities for IT. Retrieved from: http://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf

4. Wilkins, Eric. (2014, April 11). Experts urge balance between big data and privacy in health care. Princeton University. Retrieved from https://www.princeton.edu/main/news/archive/S39/72/21I52/index.xml?section=topstories

5. Gold, Ashley. (2014, April 15). Privacy could 'crash' big data if not done right. FierceHealthIT. Retrieved from http://www.fiercehealthcare.com/it/privacy-could-crash-big-data-if-not-done-right

6. Timiraos, Nick. (2016, June 22). Social Security, Medicare Face Insolvency Over 20 Years, Trustees Report. The Wall Street Journal. Retrieved from http://www.wsj.com/articles/social-security-medicare-trust-funds-face-insolvency-over-20-years-trustees-report-1466605893

7. National Health Expenditures 2014 Highlights. Centers for Medicare & Medicaid Services. Retrieved from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/highlights.pdf

8. Brown, Marie T. and Jennifer K. Bussell. Medication Adherence: WHO Cares? Mayo Clinic Proceedings. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068890/

9. CDC. Medication adherence (2013, March 27). Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/primarycare/materials/medication/docs/medication-adherence-01ccd.pdf

10. The Canadian HIV/AIDS Legal Network. (2004) HIV/AIDS and the Privacy of Health Information. Retrieved from http://www.aidslaw.ca/site/wp-content/uploads/2013/04/e-info-privacy+-+ENG.pdf

11. Health Services Research and the HIPAA Privacy Rule, National Institutes of Health. Retrieved from: https://privacyruleandresearch.nih.gov/healthservicesprivacy.asp

12. Hall, Susan D. (2016, February 17). ONC: Patient comfort levels with EHRs, data-sharing on the rise. FierceHealthIT. Retrieved from http://www.fiercehealthcare.com/it/onc-patient-comfort-levels-ehrs-data-sharing-rise

13. McKinsey Global Institute. (2011, May). Big Data: The next frontier for innovation, competition, and productivity. McKinsey & Company. Retrieved from http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation