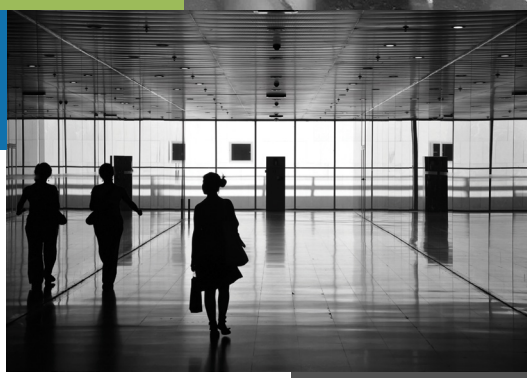
PRIVACY ANALYTICS

WHITE PAPER

Anonymization for Analytics

Data has become a precious, prized asset for healthcare organizations looking to control costs and improve patient care that can capture and action the considerable insights of big data. Statisticians and analysts are juggling not just the challenges of sifting through the growing volume, variety and velocity of data, but of safeguarding it under stringent regulatory and legal requirements. By incorporating anonymization into data management and analytic practices, organizations can establish a repeatable process for analysis into their data flows.



PRIVACY ANALYTICS

a QuintilesIMS company

Opportunity does knock, especially for organizations that can capture and action its considerable insights. McKinsey estimates that leveraging big data in healthcare - analyzing it and apply its insight to underlying processes - could account for $300 to $450 billion in reduced health-care spending, or 12 to 17 percent of the $2.6 trillion baseline in US healthcare costs.[1]

Yet, for many organizations - from payers to providers to drug companies - this seemingly straightforward opportunity belies the vast complexity of accessing and analyzing data residing in siloed, disparate systems. Indeed, achieving a full longitudinal view of patient encounters within this industry's value chain remains stubbornly elusive. Even with the emergence of health information exchanges (HIEs), the ability to analyze data from different jurisdictional sources, claims, admission and discharge date remains technically and rightly challenging legally in terms of protecting personal health information.

The implications of siloed data analyses, however, have a cascading impact on the efficacy of healthcare service delivery, patient care, drug costs and the preventive management of patient well-being - pressingly germane in helping chronically high-cost patients lead a healthful lifestyle.

Secondary use of health data presents an opportunity to overcome some of these hurdles. The term applies outside of direct health care delivery. It includes such activities as analysis, research, quality and safety measurement, public health, payment, provider certification or accreditation, marketing and other business applications. A key aspect of leveraging data for secondary use is how data is anonymized.

Anonymizing Data within an Analytics Continuum

Traditional approaches to anonymization, also referred to as de-identification, have focused on masking, primarily for application testing. But that's only half of it. Real anonymization not only masks personal information, but statistically de-identifies certain aspects of personal information using a risk- based approach that enables the application of transactional and more advanced analytics.

Critical to applying anonymization to healthcare data is how it supports organizations' efforts to be data driven, to creating and maintaining an analytic continuum that derives insight across a healthcare value chain. This continuum consists of progressive applications of business intelligence reporting, scenario modeling and predictive capabilities, ideally to a consolidated and de-identified data environment.

Stages associated with an analytic continuum could, though not exclusively so, include:

- Manual reporting from transaction systems, such as a prescription claims and refill periods.
- Healthcare scenario analytics that examine the impact of preventive care and therapeutic compliance on future health costs.
- Advanced or predictive analytics that help organizations detect fraud, or identify patient

Anonymizing Data within an Analytics Continuum

Automating Statistical De-identification

Many organizations need greater analytic utility associated with this information. For instance, all masked dates, not just date of birth, preserve a single calendar year, making it difficult for any meaningful analysis on this masked date field.

Privacy Analytics' de-identification software not only has masking capabilities, but also automates statistical de-identification. The statistical method allows for the creation of de-identified datasets that have high analytical value. Our software can de-identify quasi-identifiers/or indirect identifiers – such as date of birth, ZIP, diagnosis codes, to name a few.

The level of de-identification is based on measuring the risk of re-identification of a given dataset against a threshold determined by its planned use and legal controls in place to prevent its disclosure. This enables organizations to share or release their datasets within jurisdictional regulatory requirements cost-effectively and quickly.

and member receptivity to preventive marketing information.

Anonymization can play a critical role in each stage. It does so by giving access to data that would otherwise be hidden behind walled gardens meant to protect patients from prying eyes or inadvertent disclosures. It's a good thing these walled gardens exist, but so is getting access to high quality data that can be used to improve healthcare.

While an important part of anonymization, masked data is not useful for analysis. The point of masking is to scramble the data so nothing identifying is left behind. You could drop a column of first names, or replace them with first names selected at random from a database of fake names. Even if you randomly select female names for female patients, and male names for male patients, those masked names hold no analytic value.

Now imagine you mask date of birth. What's typically considered a crucial piece of information is rendered analytically useless. Masking provides no analytic utility. It's used for what are called direct identifiers - those elements in the data that can be used by themselves to identify individuals. You wouldn't want to use direct identifiers for data analytics, anyway.

Date of birth is actually an indirect identifier - a field that is used in combination with others in order to re-identify someone - and for this we use de-identification techniques. Indirect identifiers are used in combination with other indirect identifiers in order to re-identify someone, they are evaluated in the content of risk, of whether these data points would re-identify an individual. So date of birth, sex, and postal code would be considered together to determine who is at risk of being re-identified, if these are all the indirect identifiers in your dataset.

Now we are in the domain of analytic utility again (or at least useful analytics). Date of birth could be converted to month/year, or year, or two-year intervals, etc., and each option provides different levels of utility, and different levels of risk. It also depends on the context of the analysis. This will determine the risk threshold of a subject for another time.

Applying BI and Advanced Analytics to Anonymized Datasets

Gaining Insight from Anonymized Datasets

A leading healthcare analytics firm used Privacy Analytics to de-identify an EMR database for analytics purposes.

The EMR database covered 535,595 patients in the province of Ontario, with 5,850 providers working in 2,664 clinics across the province.

The data itself included all clinical encounters, billing information, claims (which included diagnosis and procedures), drugs prescribed, laboratory data that was sent to the EMR database for all tests ordered by the providers, and basic demographics on the patients. A total of 75 tables from the EMR database were de-identified.

As a result, a wide range of analyses can be performed, including post-marketing surveillance of drugs, public health surveillance, as well as evaluating the number of patients who meet screening criteria for clinical trials.

You can find read about this case study in more detail in *Anonymizing Health Data: Case Studies and Methods to Get You Started.*

Classic business intelligence, or BI, tools focus on queries across multiple dimensions. For instance, an analyst might want to know the number of 30-39 year old female patients in the cardiac care wing for a particular year. At the center of those queries is an analyst figuring out what's needed to run their analysis and find useful patterns.

At the other end of the value chain would be advanced analytics. In this case, the analyst could be looking to set up a predictive model to estimate the demands on the cardiac care unit every month. So it's not just a backwards glance at what the demands were, but a look at what they will be in the future.

Analytically useful de-identified data maintains referential integrity, so that relationships between data elements can be teased out of a patient's medical information. This is needed for many forms of advanced analytics - such as evidence-based medicine, predicting resource demands, or estimating patient health-risk factors.

Patient ID's, used for linking records and tables, require we maintain referential integrity. Otherwise, there would be no longitudinal de-identified data. One solution is to use Format Preserving Encryption, or FPE on any unique identifying number - internal patient ID's, medical record numbers, device ID's and serial numbers, among others.

Encryption passwords can be securely stored so that someone with authority can re-identify patients—for example, in cases of fraud detection or direct marketing, separating the analytics from patient identity. Patient ID's are, however, direct identifiers, so we are referring to masking here.

In terms of indirect identifiers and de-identification, we might randomize dates within year (i.e., following Safe Harbor) with classic BI tools. But the need for high quality anonymized data becomes even more apparent with advanced analytics. A predictive model could focus on monthly or weekly trends, be dependent on patient flows (e.g., direct admissions, transfer from other departments), or consider primary diagnoses leading up to a visit to the cardiac care unit.

By building privacy into an analytics continuum, what is called a Privacy by Design approach, organizations become compliant with privacy regulations

Maintaining the referential integrity of dates, so that order is preserved and re-identification risk is very small, is tricky. It's not enough to move all dates by some fixed amount, because this does nothing to the intervals between dates. Dialysis treatments or chemotherapy, to name but two examples, could provide a unique sequence of intervals between dates. Part of the de-identification must therefore include these intervals.

Instead, our anonymization approach takes the first date in a patient's sequence, and randomizes it within a level of generalization that is determined by the level of risk involved. It can then randomize the intervals between dates, again within some level of generalization determined by the risk involved. So, depending on the risk, you could randomize the first date to within a month, and randomize the intervals to within a week.

On average, this process of shifting the dates to maintain their order produces a sequence that maintains the length of time between a series of events for a patient. That means models looking at preventive care, for example, are still predictive because the data is more than just properly ordered, it is of high quality. When we look at the data before and after de-identification, the statistical properties are maintained.

All of this leads to a crucial point - you need to measure re-identification risk. Otherwise, how do you know what level of generalization to use? Heuristics might work in some cases, but how do you know if you've never measured the risk? And if you've only measured the risk once, how do you know it still works, or will work with different datasets? Re-identification risk needs to be below an established threshold, making qualitative assessments hard to justify in practice.

By building privacy into an analytics continuum, what is called a Privacy by Design approach, organizations become compliant with privacy regulations. But they, or their partners, also get access to highly useful anonymized data for secondary purposes.

This point can't be understated - having access to healthcare data that is compliant with privacy regulations and highly useful is the key to leveraging advanced analytics. The accuracy, and therefore usefulness, of predictive modeling is dependent on having access to quality data. Errors in the data are magnified in predictions.

Incorporating Anonymization into Your Analytic Continuum

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

Anonymization should be part and parcel of organizations' data management and analytic best practices. But just as data quality and security are critical components of data management, anonymization serves as a critical preparatory step to analyze data for secondary purposes.

In short, anonymization must take into account masking and statistical de-identification approaches to gain optimal analytic utility for secondary purposes. By incorporating anonymization into data management and analytic practices, organizations establish a repeatable process for analysis and build in critical privacy principals into their data flows.

Sources:

1 McKinsey & Company, "The 'Big Data' Revolution in Healthcare: Accelerating Value and Innovation," April 2013