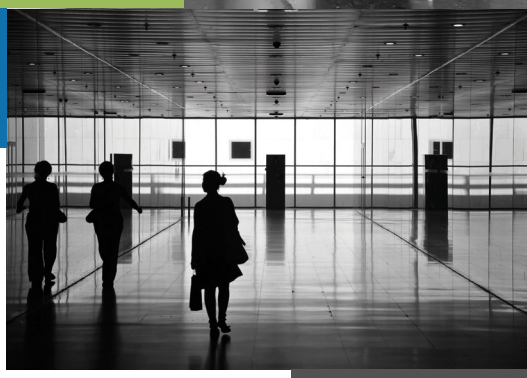
PRIVACY ANALYTICS

PRIMER

RWE & Risk: Balancing the Demand for Deep Data with Patient Privacy

The wealth of patient-level data coming in novel datasets is delivering ground-breaking insights from RWE analysis but is also driving concerns over how to ensure a patient's privacy is protected. In this primer, we describe the techniques, software platforms and highlight example use cases of how pharma companies can take both sustainable and secure approaches to access new datasets and build RWE data networks.



PRIVACY ANALYTICS

NOW PART OF IMS HEALTH

Executive Summary

As pharma companies have become more sophisticated in their use of real-world evidence (RWE), they have moved beyond standard datasets from vendors to access a deeper and wider variety of datasets, often working in partnership with local health systems. Indeed, leading pharma companies have built comprehensive networks and data platforms can that provide a shared understanding to teams across the organization about the reality of what is happening in healthcare. The increasing number and variety of datasets analyzed — including novel sources from social media through to medical imaging — are delivering ground-breaking insights.

However, accessing a growing range of data sources will necessitate new capabilities, including the critical need to protect patient privacy. Many pharma companies cite protecting privacy as one of their primary imperatives in building RWE into their capabilities, but also a key barrier to making progress. Real-world data (RWD) is patient-level data drawn from a variety of sources that all contain varying amounts of protected health information (PHI). Removal of PHI is a critical first step to using this data in RWE analysis. The challenge is how to effectively anonymize the data without diminishing data quality in an exponentially increasing number of contexts.

Fortunately, new software enabled capabilities now exist to address this urgent challenge. Best practice approaches and guidelines have emerged advocating for a risk-based approach to de- identification in order to balance the competing goals of anonymity and quality. Levering an automated de-identification process that uses a risk-based methodology ensures a continuous — and legally compliant — flow of data for RWE analysis.

In this paper, we describe the techniques, software platforms and highlight example use cases of how pharma companies can take both sustainable and secure approaches to access new datasets and build RWE data networks.

Greater Data Variety Leads to Greater Privacy Considerations

The proliferation of large-scale patient datasets, along with an increasingly competitive landscape, is placing a renewed focus on RWE. For many years, the onus has been on pharma companies to demonstrate to healthcare's payers, patients and providers the value of their medications.Now, pharma companies are facing additional challenges:

- Generic equivalents being substituted for branded drugs;
- An increased number of innovative products targeting the same therapeutic areas, and;
- Ongoing scrutiny from public and private payers that are introducing further cost containment measures and added restrictions on reimbursements.

Subsequently, innovative drug companies are increasingly challenged to maintain their strategic advantage. Those that want to keep their edge are moving beyond their reliance on standard RWD datasets from vendors to seek deeper and wider sources of data. RWD provides a view into the reality of what is actually happening in healthcare and can inform decisions that are being made across the product pipeline; from investments in development to pricing to physician and patient targeting.

RWE uses data gathered from real-life patients to support cohesive decision-making at multiple points along the drug development process. The creation of comprehensive data platforms that link information from multiple sources, such as electronic medical records (EMRs), claims databases, prescription data, lab results and social media networks, offers a 360-degree view of the patient experience. Mining this collection of data can unearth novel insights on patient populations, treatment pathways and gaps in therapy.

The need for pharma companies to access an increasing number and variety of datasets from around the world is also creating increased privacy considerations. Each of these new datasets comes with varying levels of protected health information (PHI). The removal of PHI is a critical first step that must take place prior to RWE analysis being used to inform research questions or business problems. When patient data is shared for purposes other than primary care, privacy laws require that the information is first anonymized to protect the patient.

Establishing an automated data de-identification process that uses a risk-based approach to anonymize data can shorten timelines and ensure a continuous flow of data for RWE analysis. As RWE matures, analysis will rely more heavily on environments that integrate many different sources of data. Forward-thinking organizations should aim to establish solutions that can deal with various data types and the issues that arise from linking datasets.

RWE Analysis Relies on Robust Data De-identification

Organizations investing in RWE are evolving from its use to answer questions in an ad hoc manner to applying RWE in a more systematic way across the enterprise. No longer is RWE a tool that is only applicable in a product's post-market phase to monitor drug safety and access. Industry leaders are integrating RWE into the overall product lifecycle so that every functional area of the company can use it to make better commercial decisions. RWE is starting to be used to assist in recruitment for clinical trials, improve product launches, target the right prescribers and patients, and support ongoing access through creative pricing and reimbursement mechanisms.

However, even leading companies that have established comprehensive networks and data platforms are only beginning to realize the full potential of RWE's analytic tools. An explosion in the volume and variety of RWD coming from EMRs, disease registries, genomic data, social media and wearables technology is enabling deeper and more extensive data to be integrated into the RWE platform. Investing in access to these datasets can present pharma companies with major opportunities — and also introduce major risks. Deep data contains a high degree of sensitive information that, if not dealt with correctly, can impact on a patient's right to privacy and open up an organization to regulatory violations.

To use a patient's PHI in RWE analysis, a pharma company can try to obtain the patient's consent to share their data for secondary uses. While many patients are willing to share their data for use in research, people also have a general expectation that their privacy will be maintained. As a result, anonymization in some form should be executed before using or sharing patient-level data, even if consent is obtained.

Adding Values to RWE with Deep Data

Traditionally, pharma companies have used a variety of standard datasets to help them understand how their products were working and the economic impact of their products. These sources include clinical trials datasets, claims data, prescription data and hospital administrative data. While these datasets have limitations in their analytic usefulness, they also pose a smaller risk to patient privacy.

Claims and prescription datasets contain a limited amount of patient identifying information. The discrete pieces of demographic data that could be used alone or in combination to uniquely identify a particular individual are few. Typically, basic masking techniques have been used on these datasets to provide anonymity. Masking uses either suppression to remove identifiers, randomization to replace identifying values with fake values, or pseudonymization, which creates pseudonyms in place of actual values.

The use of masking, however, introduces two issues with respect to the data: masking is detrimental to the analytic utility of data and

contains no methodology that ensures the masked data is truly anonymous. Masking all identifiers — including quasi-identifiers, like age, gender and geographic location — can destroy the utility of the dataset for analysis. Failure to mask other potentially identifying information, like dates or diagnosis codes, leaves open the possibility that unique values, or combinations of values, could be used to positively re-identify a person.

The value brought by deep data sources stems from the richness of the clinical data that they contain. Access to deep data provides two realms of opportunity for pharma companies. First, it allows analysts to tackle more challenging questions than has previously been possible and, secondly, it provides the potential to make ground-breaking discoveries that can then be acted upon; but only if these sources retain high-quality, granular data after anonymization.

The Need for Truly Anonymized Data

It is the combination of broad data (from claims and pharmacy datasets, for example) and deep data (provided in registries and specialty EMRs) that can give a comprehensive view of the patient experience — their interactions, perceptions and responses to care. It also provides functional areas across the enterprise — from R&D to HEOR and Safety to Commercial — with a common fact base about what is happening in healthcare today.

This is driving an ever-increasing demand to integrate deep data sources into current RWE environments. It can also open up a pharma company to privacy risks if PHI is not dealt with properly to ensure patient anonymity. Linking together various sources of data can increase the number of quasi-identifiers associated with each record in the dataset. The greater the number of quasi-identifiers there are, the greater the probability that there will be an individual in the dataset that has a unique combination of values for these quasi-identifiers. This uniqueness renders that person identifiable.

The risk of uniqueness, therefore, substantially impacts on the company's risk exposure. However, stripping all quasi-identifiers from the data would obliterate the data's analytic utility. Removing PHI and anonymizing the data must be balanced against data quality concerns. Simple masking solutions no longer suffice in this situation.

It is possible to have data that is truly anonymous and that still remains useful for analysis. The solution is the use of a risk-based methodology to de-identify data, like Expert Determination. This approach requires that a person with knowledge and experience in statistical methods and probabilities oversee the de-identification process to ensure that the chances of re-identifying any individual from the data is effectively zero.

With Expert Determination, information like dates can be generalized or aggregated rather than eliminated. Other techniques available with this risk-based method, like date shifting, allow chronological information and durations to be retained. All of this enables better information for use in subsequent data analysis, providing richer and more accurate analytical findings.

Expert Determination is the recommended approach to anonymizing data for RWE. Operating as they do in a global marketplace, it is prudent for pharma companies to follow the

(Risk-based de-identification) allows organizations to be confident that they are minimizing their risk exposure and operating in a manner that is compliant with privacy legislation

guidance of internationally respected experts when it comes to the use and sharing of healthcare data. Recognized industry associations including the Health Information Trust Alliance (HITRUST), the Institute of Medicine (IOM), the Canadian Council of Academies and the European-based organization PhUSE have all endorsed the use of a risk-based methodology to de-identify healthcare data.

Furthermore, Expert Determination is one of the two acceptable forms of data de-identification noted under HIPAA, the legislative foundation for healthcare data privacy in the U.S.

Establishing an Automated Flow of Anonymized Data

In addition to providing privacy compliant and granular data, Expert Determination has the added advantage that data de-identification can be automated since it is based on statistical principles and methods. Establishing an automated data de-identification system is the cornerstone to a robust RWE environment.

When data is being continuously updated, the use of an automated process helps to apply data de-identification consistently to incoming data.

Pharma companies that use RWE across the enterprise need to regularly incorporate new data into their networks so that they are accessing the most currently available information. The implementation of an automated data de-identification system gives analysts and researchers timely access to a continuous flow of current data in an anonymized format, a situation that would be nearly impossible using manual processes.

Implementing an Automated Data De-identification System

As with any risk-based approach to de-identification, the first step is to assess the privacy risk. This depends on the context of the situation — who will have access to the data, what security and privacy controls are in place to protect it from unauthorized access and how sensitive is the information. Assessing the context is critical; it will be used to determine how much de-identification is needed. Should dates be grouped by month or by year? Are there unique values that need to be suppressed? Without

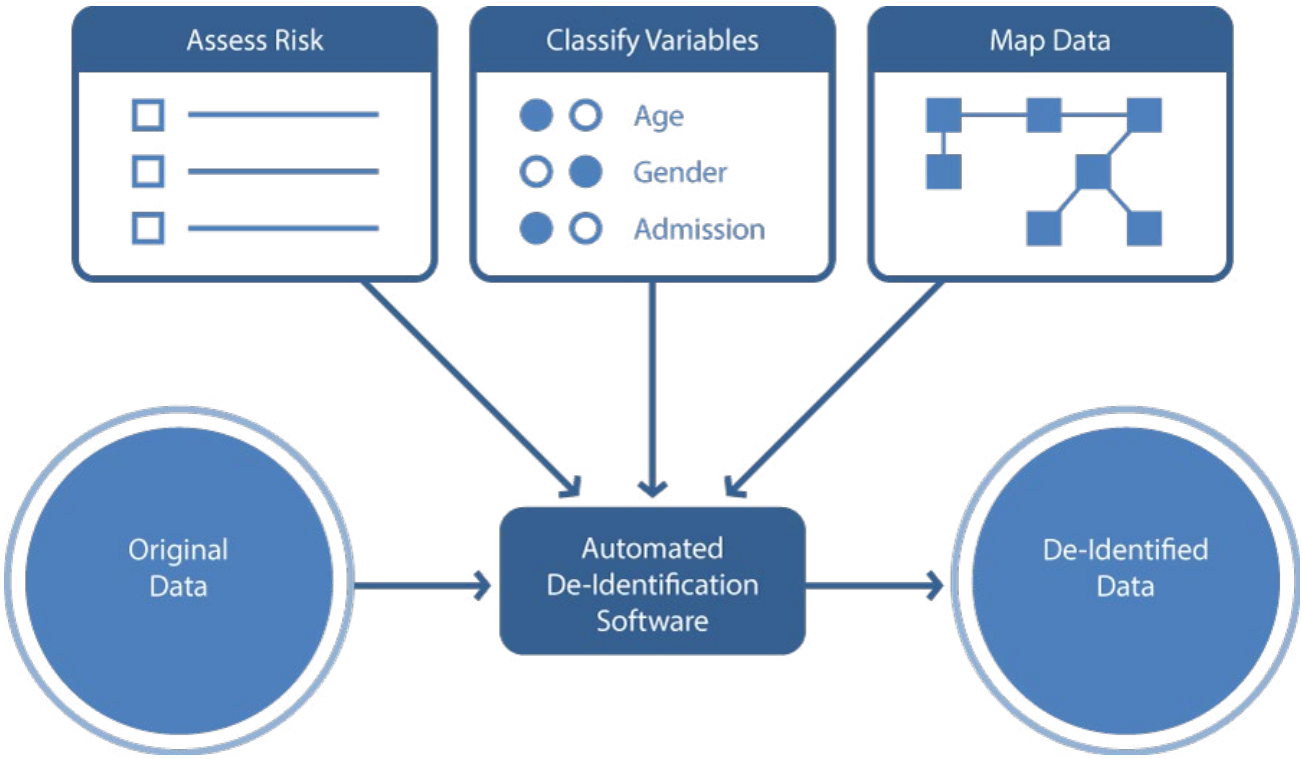


Diagram 1: Schematic of an automated data de-identification pipeline.

knowing the context — and the variables that are important to the subsequent analysis — it will be impossible to find the correct balance between data anonymity and data quality.

Once the context is assessed, the next step is to classify the variables in the data. While a dataset may contain hundreds of tables with thousands of variables, only a subset of them contain keys to an individual's identity and are relevant from a privacy perspective. These are the variables that we focus on for de-identification. They are classed as either direct identifiers (for example, name or social security number) or quasi-identifiers (such as age, birthdate or profession). Direct identifiers are suppressed or pseudonymized since these variables alone can be used to identify a person. Since direct identifiers have little use for analysis the removal of real values does not pose an issue. The quasi-identifiers is where the bulk of the effort is directed. It is these fields that are impacted by the assessment undertaken in step one and where we'll need to make adjustments to turn up or turn down the level of data manipulation.

The final step is to map the data. This is a technical step that ensures the de-identified data maintains the integrity of the original database. To confirm the integrity of de-identified healthcare data, tests have been run that compare the results of a research protocol that first used an original dataset containing PHI, followed by the use of the same dataset after de-identification.

The test, which looked at the prevalence of gastrointestinal adverse events in patients taking NSAIDs both with and without proton pump inhibitors, showed that the de-identified data delivered results that were consistent with the original dataset and that led to the same conclusions.

Delivering Consistently Low-Risk Data

With the implementation phase complete, the real work of the of the automated data de-identification system can begin. This process permits data to be pulled in on a regular and recurring basis (e.g., weekly, monthly or quarterly) from the data source and exported to the pharma company's RWE platform.

To limit the risk from a re-identification attack, the RWE platform only ever accepts data that has been de-identified. Data de-identification is performed at the source site with the automated de-identification engine performing the necessary steps to remove or perturb the identifying variables. The resulting dataset is then measured. This is to confirm that the level of risk falls below the acceptable risk threshold for each cut of the data that will be exported. If the risk level does not fall below the threshold, the dials are adjusted to further manipulate the data until this is achieved. Once the risk level is sufficiently low, the de-identified data is then exported to the RWE data warehouse where analysis can be run.

Establishing an automated de-identification system lets pharma companies quickly and consistently de-identify millions of patient records when refreshing the content of the data warehouse for RWE. It also allows organizations to be confident that they are minimizing their risk exposure and operating in a manner that is compliant with privacy legislation.

By engaging with experts in the field of de-identification, an automated de-identification system can be quickly and efficiently implemented that is in line with privacy legislation, like the HIPAA Privacy Rule. In the event of a data breach, the ability to show practices that comply with the legislation provides companies with a defensible position in the event of a lawsuit.

Conclusion

The widespread use of RWE to inform product pipeline decisions in the pharmaceutical industry is not yet the norm. However, leading companies are actively seeking novel sources of deep data to spark ground-breaking insights and gain a competitive edge. A leading information and technology services company has estimated that, by applying RWE in a systematic way, a top-ten pharma company could realize $1 billion in value. However, increasing the number and variety datasets incorporated into data networks for RWE analysis can increase a pharma company's exposure to risk. Deep data sources provide more extensive patient- level data that — while providing enormous opportunities for future product development — necessitate more sophisticated approaches than simple masking solutions to address privacy concerns. Data anonymity cannot be delivered at the sacrifice of data quality or else RWE analysis could be rendered pointless.

Risk-based data de-identification methodologies, like Expert Determination, allow for a balance between data anonymity and data quality, with consideration for how the data will be used. With the number of contexts for the application of

CONTACT US

251 Laurier Ave W
Suite 200, Ottawa, Ontario
Canada K1P5J56

Phone: 613.369.4313

www.privacy-analytics.com

sales@privacy-analytics.com

RWE increasing, determining the data's context for use is the foundation for maximizing the value obtained from the data. Use of a software-based automated de- identification process can provide high-quality, granular data that is tuned to optimize utility while definitively adhering to privacy requirements under HIPAA, PHIPA and the EU's General Data Protection Regulation.

Learn more about RWE and applying risk-based de-identification. Read our case study with IMS Brogan: http://www.privacy-analytics. com/files/IMS-Brogan-Case-Study.pdf.

Sources

1. For more on Expert Determination, see Privacy Analytics' white paper *De-identification 201: Fundamental of Data De-identification* available at http://www.privacy-analytics.com/de-id-university/white-papers/de-identification-201/.

2. Hughes, Benjamin, Marla Kessler and Amanda McDonnell (2014). Breaking New Ground with RWE: How Some Pharmacos are Poised to Realize a $1 Billion Opportunity. IMS Health. Retrieved from http://www.imshealth.com/files/web/Global/Services/ Services%20TL/rwes_breaking_new_ground_d10.pdf.