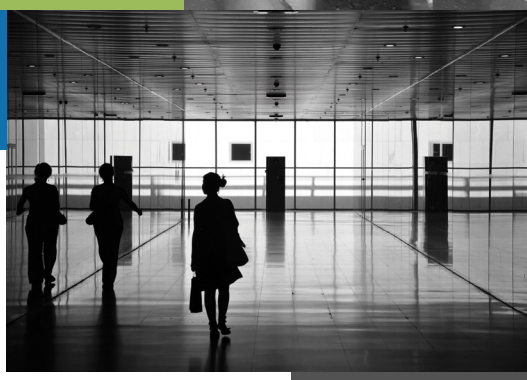
PRIVACY ANALYTICS

WHITE PAPER

Strengthening Secondary Use: How Registries Can Responsibly Share Data

Balancing privacy, risk and data quality has become crucial for academic medical centers. As the demand for data increases so do the privacy risks around linking multiple data sources. As the need for registries grows, evidence highlights that statistical de-identification is the optimal choice to deliver high-quality, anonymized patient data. This white paper outlines how to establish best practices for combining both.



PRIVACY ANALYTICS

a QuintilesIMS company

Executive Summary

Data registries are an invaluable tool that require the secondary use of data for conducting evidence-based healthcare research. Concern for patient confidentiality has always been a factor when it comes to sharing data. Currently, the growing prevalence of data linking between repositories of patient information is heightening risks to privacy and driving concern to a new level.

In order to provide comprehensive, detailed data on specific patient populations, disease registries will link patient information from electronic medical records (EMRs) with claims data and other administrative files. While these linkages provide a more complete picture of the patient experience, they also associate a greater number of direct and indirect identifiers with an individual patient record. The greater the amount of data that is available on an individual, the greater the chance that they could be re-identified.

Anonymizing patient data to remove protected health information (PHI) is essential before information from a disease registry can be disclosed. However, when data is being shared for research and analysis it is also important that it retain its analytic utility. Statistical de-identification is the only method that allows data quality to be maintained while ensuring that the data is truly anonymized. Although other methods exist to remove PHI, statistical de-identification is the optimal method to anonymize patient data, particularly when it is to be shared for secondary use.

In discussing how registries can responsibly share their data with researchers, this white paper will review the option of patient consent versus de-identification and explain how statistical de-identification can provide researchers with the highest-possible quality data for their needs.

Balancing the Benefits and Risks of Disease Registry Data

A staggering amount of data is collected on patients nowadays. Turning this data into knowledge, however, requires that it be available and accessible in ways that support healthcare research. Disease registries have become an invaluable tool, enabling researchers and analysts to gain tremendous insights on specific patient populations.

All of this research requires that data from disease registries gets shared responsibly — or else patient privacy could be compromised. Registries gather information about people who have a specific disease or condition or who have undergone specific interventions or procedures. Patients that have a stigmatizing health condition may be left vulnerable to social, economic or psychological harms if their privacy is breached. These databases can tie together information ranging from a patient's demographics (race, gender, marital status, etc.) to income and education levels to prescription information and lab test results. It is information that is highly personal and often highly sensitive. Ensuring the confidentiality of the data contained in these registries is of paramount concern.

Data linking is being seen more often as registry owners look to increase the value to their holdings by integrating data into a single repository. The abundance of data that a registry gathers on patients with a given health condition (or who have undergone a certain procedure) can provide substantial opportunities for secondary use. Registries have been used to identify candidates for clinical trials, support ongoing disease surveillance, inform healthcare policies and enrich our understanding of the causes, patterns and progression of many diseases. The ability to look at entire populations of patients to glean patterns and trends is opening up new pathways in the prevention and treatment of disease based on the available evidence. This is driving huge demand for access to these rich and detailed datasets.

While the benefits of sharing data from registries are many, this must be offset against the potential risk to patient privacy. It means the need to find the proper balance between access to data and anonymity. Total access would almost certainly allow for patients to be positively identified, while total anonymity would render the data significantly less useful. Fortunately, methods exist that allow high-quality data to be shared securely.

Data de-identification protects patient confidentiality by effectively removing protected health information (PHI) from the data. The use of a risk-based approach to de-identify data, like HIPAA's Expert Determination, is recommended by leading data organizations the world over. Risk-based de-identification allows data to be made anonymous while still retaining its quality for research purposes. It is the only approach that allows registries to responsibly share their data.

Better Access to Data Requires a Better Understanding of Privacy Risk

Historically, accessing healthcare data for research was difficult. Before the shift from paper charts to EMRs, the available data was limited. Accessing it often meant having to physically go to the location where the data was stored. This presented barriers to research since only a small number of investigators could work with the data at any one time. Research done this way was also costly since data owners would need to work with researchers to identify and gather the necessary data. These factors put constraints on the amount of research that happened, but also limited the risks to patient privacy.

The widespread use of EMRs has driven the growth in healthcare data and has led to a concomitant rise in the creation of patient registries. Registries have been developed by federal and state governments, professional organizations, patient advocacy groups, and hospitals, covering everything from the most common chronic conditions like diabetes to rare genetic disorders. While some of these rare disease registries may contain information on a very small portion of the population, other mega-registries like the SEER-Medicare database link together sources of data from multiple states to provide detailed information about Medicare beneficiaries who have cancer.

The increase in digital data and the ease with which is can be accessed via VPNs or the internet has made information easier to share, but this brings with it greater obligations on the part of registry owners to minimize the potential harms to people whose information resides in the registry. Data is more easily accessed but patient privacy is more easily compromised in this environment. The size and scope of these datasets, the escalation in requests for data, and the cultural changes in our perceptions and expectations around privacy are all placing new requirements on registry owners to be aware of privacy issues and to understand what actions they need to take to reduce their susceptibility to a data breach.

Virtually all of the data that is collected in registries comes from sources that are considered covered entities or business associates under HIPAA. The HIPAA Privacy Rule sets out the standards for the use and disclosure of PHI held by covered entities and their business associates. As such, complying with HIPAA helps registry owners mitigate their privacy risk.

Under HIPAA, it is possible to use a patient's PHI for secondary uses like research if they have provided their consent for such use. However, obtaining patient consent can be cumbersome and present problems. The preferred solution is to de-identify the data in a manner that is compliant with the standards set out in the Privacy Rule. Of the two de-identification methods outlined, the statistical method — also known as Expert Determination — provides distinct advantages for both data quality and privacy.

While de-identifying data will not stop an attack or prevent a data breach from occurring, it can

make it effectively impossible for the attacker to positively re-identify someone from their data or to get ahold of information that would be useful for nefarious purposes. In this way, it provides an added element of security in the event that data is compromised.

Consent, De-identification and the HIPAA Privacy Rule

Consent versus De-identification

Most patients want to support healthcare research and are not opposed to their health information being used for this reason[1], but they do retain an expectation of privacy even in these situations.

Like most privacy laws around the world, HIPAA is consent based. As such, a patient's PHI can be disclosed for use in health research or other secondary purposes provided that he or she has granted authorization for that use. While it may seem reasonable to try to obtain consent from the patient when data is collected, it would be impossible to inform them of all of the possible future uses of their data. Thus, the difficulty comes in obtaining informed consent.

The alternative is to get consent after the fact, once a specific research problem or use for the data has been identified. This also introduces problems of practicality. Contacting the millions of individuals who have data held in the SEER-Medicare database, for example, would be an expensive and time-consuming process. Ultimately, it would prove to be a futile activity since some patients will have moved, changed their contact information or died.

Even if all of these hurdles can be dealt with, obtaining consent can result in bias in the data which has negative consequences for data quality. It has been shown that consent requirements lead to an ascertainment bias[2]. Individuals who consent to the use of their data tend to have different characteristics than those who do not consent. By reducing participation in research on the part of some groups of individuals, the research sample is non-random and, therefore, does not reflect the entire population of the health condition in question.

Thus, even if consent could be obtained, it is advisable to de-identify data instead. By stripping the PHI from a dataset, it is possible to share data while avoiding consent bias and providing assurance to patients that steps have been taken to protect their privacy.

The HIPAA Privacy Rule: Safe Harbor vs Expert Determination

If the PHI is removed from registry data, then it is no longer subject to HIPAA restrictions. HIPAA's Privacy Rule provides mechanisms for using and disclosing health data responsibly without the need for patient consent. To guide covered entities and business associates on how to eliminate PHI from their data, the Privacy Rule specifies two standards that may be used to de-identify data: the Safe Harbor method and Expert Determination.

Safe Harbor is an easy-to-follow, prescriptive approach to de-identification. In this method, 18 data elements are listed that must be removed from the data or generalized. The list includes names, phone numbers, email addresses, social security numbers and all elements of dates except for the year, among others. Safe Harbor

does not concern itself with the subsequent quality of the de-identified data, taking the same approach regardless of the data's context for use or the research requirements. So, if a researcher wishes to analyze data for seasonal variations in acute respiratory cases and needs to know the month of hospital admission for this research, the information cannot be provided using Safe Harbor; only the year can be provided for dates.

The other approach to de-identification that can be used is Expert Determination. This method takes a risk-based approach to de-identification based on statistical principles. In other words, it measures the risk that a person could be re-identified from the data and then, as necessary, uses best practices to perturb the data in order to adequately mitigate that risk. In this approach, context is taken into account with consideration for the security controls the data recipient has in place, the sensitivity of the data, and the motives for re-identification. This allows the right balance between data access and patient anonymity to be found. The correct amount of de-identification can be applied to the data — no more and no less — to achieve a risk of re-identification that is effectively zero while retaining the highest level of data granularity possible.

For registries that need to share data that is analytically useful for research purposes, the use of Expert Determination is the only truly viable option for de-identification. This approach to sharing data not only helps to avoid issues of consent bias but also allows for flexibility in how anonymity is achieved. This flexibility is an important factor when responding to multiple data requests that each have unique requirements from the data. Expert Determination is also the approach to de-identification that is consistent with recommendations set out by leading data organizations around the world like HITRUST, the Institute of Medicine, PhUSE, and the Canadian Council of Academies.

The use of Expert Determination requires that a person with appropriate knowledge of, and experience with, generally accepted statistical and scientific principles and methods render the information not individually identifiable. However, software solutions are available that, once implemented, semi-automate the de-identification process based on these expert principles.

Expert Determination Allows Trade-Offs To Be Made Between Data Specificity and Anonymity

Unlike criminals who try to profit from illegally obtained data, researchers who study the health of populations rarely have any direct interest in knowing the specific identities of the people they study. Their focus is on aggregate trends. Thus the removal of direct identifiers, like name, through data de-identification is not a concern from a health research perspective. It is the removal of other information that can potentially identify individuals — the quasi-identifiers — that presents the issue. Information like gender, age, profession, postal codes and dates are often useful for research but can be used in combination with one another to hone in on a unique individual.

In protecting the confidentiality of patients, registries may need to consider how they source data as well as how they share it. For registries that obtain data from many different sites in multiple states, it may be beneficial to de-identify

the data at the source, before it is brought into the registry. This can avoid complications from applicable state laws. However, many other registries will contain PHI, particularly those that are held by academic medical centers, hospitals and state governments. These registries tend to source data strictly from within their own jurisdiction.

Increasingly, disease registries are making linkages with other sources of patient information, connecting their data with other administrative and claims data, in order to create a more thorough picture of the patient experience. In the future, this data linking will only increase as interest in the potential learning from the variety of data grows. The possibility exists now to integrate genomic data, content from wearables and social media data into registries.

Even if data has previously had de-identification applied, linking data will expand the amount of information associated with each patient and can increase privacy risk. Thus, in all cases, registries should be assessing the data for the potential to re-identify individuals before it is shared with outside researchers or analysts.

In supporting healthcare research, registries are faced with requests for data that each have their own specific requirements. In fact, every data request will be unique to the objectives of the research. For example, one investigator may be interested in knowing the specific age of a patient but isn't as concerned about where in the country the patient lives. Another investigator may be focused on regional variations in disease patterns but does not mind having age data grouped into 5-year or 10-year bands. This highlights the fact that an approach to de-identification is needed that offers flexibility in anonymizing data. Delivering high-quality data quality is extremely important for research but data quality can mean different things to different researchers.

A risk-based approach to de-identification, like Expert Determination, allows data to be effectively anonymized in both of these cases while permitting more granularity for the variables that are important to each investigator. It is a

Data linking will only increase as interest in the potential learning from the variety of data grows. The possibility exists now to integrate genomic data, content from wearables and social media data into registries.

matter of making trade-offs on what data elements retain high granularity versus what is sacrificed.

Since not every data request is dealt with the same way, negotiations with the researcher are a key part of the process. Once a research request is made, the researcher would need to work with the data provider to prioritize the data elements that are necessary to the study. Only a portion of the data elements in any registry are critical from a privacy standpoint. The prioritization exercise is focused on this group of quasi-identifiers so that de-identification can be applied appropriately.

Measuring the risk of re-identification in the dataset and adjusting how much de-identification is applied to each element then becomes an iterative process. Adjustments are made to the data until the potential for re-identification falls below an acceptable threshold of risk. This is where prioritization helps. If re-identification risk is too high upon initial assessment, and the investigator has identified the need for specificity around age, then the data analyst can try increasing the de-identification on another element like zip code by masking it to a 3-digit or 2-digit code. The data would then be re-assessed to see if the risk threshold had been reached. Only when the risk is found to be acceptably low can the data be exported for use by the researcher.

It is evident that manual manipulation can quickly become cumbersome and unmanageable. Implementing software to partially automate this process can help to meet mounting data requests and allow important research to go forward more quickly.

Conclusion

There has never been a more exciting or opportune time for healthcare researchers. The increase in the availability of data, coupled with ease of being able to access data over digital networks, is opening up exciting areas of investigation into population health and personalized medicine.

It is also driving a greater focus on issues around patient privacy. The growth in the amount of information captured and the propensity for linking together multiple sources of data is creating rich and comprehensive disease registries. These registries are strengthening healthcare research and enabling new discoveries into the pathways and treatment of disease but it is imperative that data is shared with researchers in a responsible and compliant manner.

The use of risk-based de-identification, like Expert Determination, allows high-quality data to be shared for healthcare research. This approach not only protects patient confidentiality but also allows flexibility is how anonymity is achieved and avoids problems of bias that are associated with seeking patient consent for secondary data use.

Privacy will continue to be a focus going forward as advancements in healthcare bring new data into the mix. The availability of genomic data is becoming more prevalent and there is enormous research potential if this data is integrated into a disease registry. With the theoretical possibility of completely identifying someone based on their DNA, how this is addressed from a legal standpoint — and the subsequent implications for data privacy — remains to be seen.

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

Learn more about the role of risk-based de-identification in Health Registries. Learn how BORN Ontario balances patient privacy with quality, insightful data in this case study: http://www.privacy-analytics.com/files/BORN-Case-Study-2016.pdf

Sources

1. Hall, Susan D. (2016, Feb. 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.

2. El Emam, Khaled et al. (2009). A globally Optimal k-Anonymity Method for the De-Identification of Health Data. Journal of the American Medical Informatics Association. 16(5): 670-82.