PRIVACY ANALYTICS
an IQVIA company

SANOFI
CASE STUDY

A Scalable, Risk-based Approach to De-identifying Clinical Trial Data: A Case Study with Sanofi

OVERVIEW

Sanofi has a number of initiatives to share clinical trial data with the research community for secondary purposes. Maximizing the utility of clinical trial participant data provides opportunities to conduct further research, help advance medical science and improve patient care.

Sanofi is not alone in this drive to share clinical trial data. There is a growing movement to increase the sharing of clinical trial data by researchers and manufacturers. As manufacturers start to make individual participant data (IPD) from their clinical trials available, privacy concerns must be addressed.

There is a recognition within the research community that secondary analysis of clinical trial data may provide new insights compared to the original publications and analyses. In addition, the increasing complexity of science requires multiple teams to collaborate and analyze the same dataset from different perspectives, pool datasets, and link multiple datasets.

This case study describes a de-identification project conducted with Sanofi Pasteur and Sanofi to de-identify a vaccine clinical trial dataset using the Privacy Analytics methodology and software, to compare its results with Sanofi's current process, and inform Sanofi's de-identification methodology.

Privacy Considerations

In July 2013, manufacturers committed to the Pharmaceutical Research and Manufacturers of America (PhRMA) and European Federation of Pharmaceutical Industries and Associations (EFPIA) principles for "Responsible Clinical Trial Data Sharing." Since 2014, they have been sharing information with researchers through company specific platforms and multi-company platforms.

However, it is necessary to de-identify these datasets before they are shared with researchers. Sharing patient level information of research participants for secondary purposes beyond the objectives of the original trial requires de-identification or other methods to protect the privacy of the research subjects.
De-identification methods consistent with contemporary standards should be used since there is evidence that improperly de-identified data can be, and has been, re-identified successfully.

If a data custodian does not de-identify data properly, the organization is at legal, financial and reputational risk associated with re-identification. There is legal risk for non-compliance with relevant regulations, such as the HIPAA Privacy Rule in the US[1] and the EU Global Data Protection Regulation, financial risk from fines and lawsuits due to a data breach, and the reputational risk of losing patient trust when they learn that personal or protected health information was shared inappropriately.

Data Utility and Scalability

While privacy is an important concern, the clinical trial data shared for secondary purposes must also have the greatest level of utility for its intended purpose. This means the conclusions resulting from the de-identified data needs to be consistent with the results that could have been obtained from the original identifiable data.

In this collaborative project between Sanofi and Privacy Analytics, Privacy Analytics de-identified the data from the Fluzone vaccine clinical trial. Applying Privacy Analytics' software, methodologies and expertise, an acceptable level of risk for Sanofi's Fluzone clinical trial data was determined and appropriate data perturbations were performed. Comparisons between Privacy Analytics, Sanofi and other industry approaches were also conducted.

"Clinical trial data will allow researchers to unlock more insights in treatment and care than ever before. In order to share this data, we must understand the risk of re-identification and ensure we use methods that are demonstrated to protect clinical trial participants' privacy."

- Robin Jenkins, Senior Director Program Management, Clinical Trial Data Sharing, Sanofi

The Data: Fluzone Clinical Trial

The data for this comparison was generated from the clinical trial: "Immunogenicity and Safety Trial of Quadrivalent Influenza Vaccine Administered by Intradermal Route in Adult Subjects Aged 18 Through 64 Years" (clinicaltrials.gov identifier NCT01712984). The purpose of this trial was to demonstrate the safety and immunogenicity of one vaccine compared to two others in protecting against four strains of influenza virus.

A critical component to protecting privacy is understanding the context in which the data will be shared. Sanofi uses multiple platforms and mechanisms to share clinical trial data, such as the clinical study data request (CSDR)[2] platform. Sanofi makes data from clinical trials available through CSDR to researchers with research proposals that have been reviewed by an independent review panel. Only research proposals studying the same study drug or disease studied in the original trial can be analyzed. The data release is subject to certain criteria being met, including a requirement to effectively de-identify the data.

Re-identification Risk Measurement

A risk-based de-identification methodology identifies an acceptable risk threshold for the data to be shared, and evaluates the data to determine whether or not the acceptable risk threshold is met. If it is not met, the data needs to be de-identified to meet the risk threshold. This report does not cover the exact formulas used, but will explain some of the key components that determine the level of risk in the dataset. The exact formulas and calculations used for this case study have been discussed elsewhere.[3]

This methodology analyzes three plausible attacks on data: deliberate attempt, inadvertent attempt, and data breach. This approach follows standards outlined by the Institute of Medicine report on Sharing Clinical Trial Data and the EU PhUSE standard on de-identification.

These plausible attacks cover the universe of attacks that the data disclosure needs to protect against.

- Deliberate attempt - the data user or recipient deliberately attempts to re-identify the dataset.
- Inadvertent attempt - when the data user or recipient inadvertently or "spontaneously" recognizes someone in the data. This is not a deliberate re-identification attack, but an inadvertent one. It occurs when a research analyst working with or inspecting the data inadvertently recognizes someone they know.
- Data breach – an attack on the security system housing the data. If there is a data breach, then the data can effectively be attacked by anyone.

Risk-based, statistical approaches to de-identification also analyze the data itself in order to measure overall probability of re-identification. In this case study we measured the average risk in the data and the uniqueness of the participants in the data. These two metrics provide quantitative coverage of the relevant risks that need to be measured.

Re-identification risk can come from direct identifiers and quasi-identifiers. Information such as location, race and age are known as quasi-identifiers. Subject IDs are known as direct identifiers. A risk-based approach will de-identify both the direct and quasi- identifiers to minimize risk in the dataset.

The Results

Privacy Analytics calculated the acceptable risk threshold. This study found that the acceptable risk threshold for the Fluzone clinical trial data released through a mechanism such as CSDR was 0.075. This threshold is consistent with precedents for sharing health information, and reflects the fact that this is not a public data release (i.e., the data would be shared through a secure portal with the researchers signing data use agreements).

Privacy Analytics' software and methodology were then applied to the dataset. De-identification techniques were applied to information such as date of birth, age, and race to reduce risk. Certain sensitive medical information that appeared in medical history was removed from the de-identified dataset. This reduced the risk of harm that may fall on the participants if there is any re-identification. Privacy Analytics' TEXT software was used to redact sensitive information within free form text fields. This further protected the data, while increasing the potential analytical quality.

The original dataset had an overall risk of re-identification estimated to be 0.14, which is higher than the threshold. By applying the recommended de-identification methods the final risk of re-identification was reduced to 0.06, which was below the threshold.

When we de-identified the same Fluzone dataset using another common methodology described on the CSDR site we found that the risk of re-identification that was measured in the data to be 0.082, and this was higher than the threshold at 0.075. Similarly, the application of a methodology provided by TransCelerate Biopharma for the de-identification of IPD resulted in a measured risk of re-identification of 0.082, again this was higher than the threshold of 0.075. Both of these methodologies stipulate the application of a set of fixed de-identification rules on all trial datasets.

This highlights the importance of performing a residual risk analysis after applying rule-based de-identification techniques. As this example illustrates, the rules by themselves may not be sufficient to reduce the risk of re-identification to acceptable levels, even for a trial that has a relatively large participant pool and that does not pertain to a rare disease.

Company Info

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

Those methodologies would have also removed all text information, whereas in this case study we were able to retain much of the information in the text narratives. The retention of free-form text commentary is another benefit of applying a risk-based de-identification methodology, in that data perturbations can be more precisely calibrated.

"Conducting this risk assessment on our data provides additional assurances that we do not inadvertently expose personal patient information. The results of this case study will inform our internal de-identification processes."

- Paul Susheel, Manager, Statistical Programming, Sanofi Pasteur

Conclusion

Privacy Analytics' automated process produced a dataset with a demonstrated risk of re-identification that is considered very small, and that is consistent with contemporary standards and best practices. Based on feedback from the analysts most familiar with the trial and its data, the de-identified dataset retains sufficient utility for replicating the original analysis and to facilitate other innovative uses of the data.

With Privacy Analytics' software, the overall process took just under a week to complete. This was based on an analyst starting from zero knowledge about the trial itself. It is estimated that an analyst with pre-existing background knowledge about the trial would be able to complete the de-identification, including producing a risk analysis report, in half a week.

Other companies looking to share clinical trial data must ensure participants' personal information is protected at the lowest level of risk, while also allowing for the greatest potential for analysis for researchers. Organizations concerned about the amount of time it takes to analyze the risk and properly de-identify the data may consider software solutions that automate this process.

Notes

1. For Covered Entities and Business Associates.
2. http://www.clinicalstudydatarequest.com
3. http://www.privacy-analytics.com/de-id-university/webinars/de-identifying-a-clinical-trial-data-set/