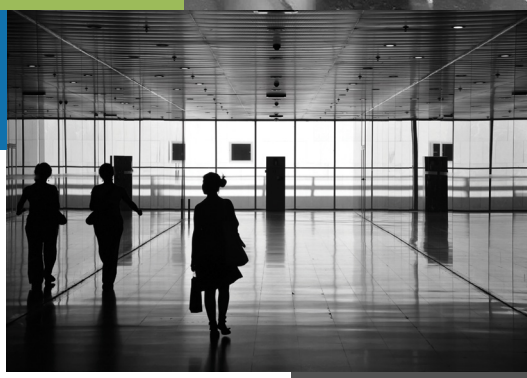
PRIVACY ANALYTICS

WHITE PAPER

Achieving Transparency in Clinical Trials

Attaining true transparency in drug trials calls for a better solution than excessive redaction. Learn the drawbacks and benefits of the different methods for anonymization in clinical trials. There is potential for strong industry leaders to emerge — so long as they align themselves with best practice guidelines. By making the right choices now, pharmaceutical organizations will meet the new transparency mandates while complying with privacy regulations and maximizing the quality of data that can be shared for secondary purposes.



PRIVACY ANALYTICS

a QuintilesIMS company

Executive Summary

Recent policy changes by Europe's regulatory authority for drugs and devices has rekindled the pharmaceutical industry's focus on clinical trial transparency. While there have been many initiatives to enhance openness in drug trials, the European Medicine Agency's (EMA) requirement to make a trial's clinical study report (CSR) publicly available is seen as a revolutionary move to boost public trust in the drug approvals process.

As the comprehensive report that details the protocol and results of a clinical trial, the CSR can contain highly sensitive health information about the study's participants — placing the issue of patient privacy front and center. It also leaves many in the biopharmaceutical community wondering how to meet the EMA's new transparency requirement while remaining compliant with privacy legislation and protecting their own confidential information.

How biopharmaceutical companies choose to anonymize their clinical trial data and reports have serious implications for transparency and their ability to leverage data for secondary uses. On the surface, different approaches to protecting the privacy of trial participants may appear to be equally sound, but they can have vastly different results when it comes to the usefulness of the anonymized content for subsequent analysis.

Understanding the drawbacks and benefits of the different methods for anonymization can help pharma companies better position themselves for the future. By aligning with best practice guidelines, they will not only be able to meet new transparency mandates but do so while complying with privacy regulations and maximizing the quality of data that can be shared for secondary purposes.

Introduction

Initiatives to improve clinical trial transparency have historically focused on registering clinical trials and publishing summary trial results. Now, recent actions taken by the EMA aim to radically improve transparency in the drug approvals process.

Last year the EMA implemented policy 0070, which requires publication of a drug trial's clinical study report (CSR) if that drug receives market authorization. Making the anonymized CSR publicly available on the web is seen as a significant step on the path to transparency. The document provides extensive details on the clinical trial, including the study objective, the investigational plan and study design, the evaluation and analysis done, and specifics about the patients who participated.

It is this last item in particular — detailed information about the experience of the study's participants in the trial — that creates huge privacy concerns and has the biopharmaceutical community wondering how to meet the EMA's new transparency initiative while remaining compliant with privacy legislation. Publishing the CSR, or sharing the individual participant data (IPD) that is at the heart of any clinical trial, introduces a risk to the privacy of the trial's participants.

While most patients support the use of their data for the betterment of patient care, they want assurance that their privacy is maintained. This has resulted in an emphasis on how to best safeguard patient privacy as it pertains to clinical trials. Biopharmaceutical industry groups, as well as organizations like the Pharmaceutical Users Software Exchange (PhUSE) and the Institute of Medicine (IOM), have published guidelines on how to minimize the risk of re-identification when sharing clinical trial data. However, some of these methods offer a more sound approach than others.

Clinical trial transparency is coming to a critical juncture. External forces pressing for greater transparency are meeting with internal drivers to leverage data for secondary uses. Both require the ability to deliver high-quality, de-identified data. Where once the conversation focused on whether clinical trial data should be shared, it is now focused on when data should be shared and how to do it. The way that study sponsors choose to anonymize their clinical trial data has serious implications for both transparency initiatives and secondary use.

This paper looks at the history of clinical trials transparency, the mechanisms for sharing trial data, and the challenges in making this data available. It then delves into the standards and guidelines for disclosing clinical trial data to see how the implementation of the different approaches can impact the data's usability.

A Brief History of Clinical Trials Transparency

Although the EMA initiative to publish CSRs came into force only recently, many biopharmaceutical companies have a deep history of making clinical trial information available on the web. It has been nearly 20 years since the Food and Drug Administration Modernization Act of 1997 (FDAMA) mandated the registration of publicly and privately funded clinical trials of human participants.

As a result of FDAMA, the National Library of Medicine (an institute of the National Institutes of Health) launched ClinicalTrials.gov in February of 2000. Since its inception, more than two hundred thousand studies have been registered on the website, a number that has more than doubled in the past five years[1]. In 2007, Congress expanded the requirements for clinical trials under the Food and Drug Administration Amendments Act (FDAAA) to add the submission of summary results to the site, including adverse events. The specifics of the submission process for clinical trial results are just now being determined.

In November 2014, the U.S. Department of Health and Human Services (HHS) issued a Notice of Proposed Rulemaking describing the proposed requirements and procedures to register and report summary trial results on ClinicalTrials.gov. These proposed changes would also expand the scope of clinical trials required to submit summary results to cover unapproved, unlicensed and uncleared products. At the same time, the National Institutes of Health (NIH) proposed a policy that would expect all NIH-funded clinical trials to register and submit summary results to ClinicalTrials.gov.

The EMA's policy 0070 is the latest attempt by a regulatory body to answer the call for greater transparency in clinical trials. As Europe's regulatory authority for medicinal products, the EMA has roughly the same function as the U.S.'s

Food and Drug Administration (FDA). Under the first phase of Policy 0070, which came into force on January 1, 2015, the agency proactively publishes an anonymized version of the CSR that it receives from biopharmaceutical companies as part of a marketing authorization application for human medicines. In addition to the CSR, companies must also provide an accompanying risk analysis report that describes the de-identification methods used and their impact on data quality.

The need to provide an anonymized public version of the CSR is causing biopharmaceutical companies to re-examine their data anonymization processes. Companies need a solution that provides a consistent and scalable approach to de-identifying both the CSR and IPD.

Mechanisms for Sharing Clinical Trial Data

Many biopharmaceutical companies embraced the idea of openness in clinical trials and undertook their own data sharing activities. A number of projects and websites have been created by industry and research groups to provide a centralized source for clinical trial information. These include OpenTrials, Clinical Study Data Request (CSDR) and Project Data Sphere. As of early 2016, data from more than 2800 trials is available on collaborative portals[2].

In addition to supporting transparency, making clinical trial data available facilitates its use for secondary purposes, which provide numerous benefits. It supports the work of analysts and researchers to advance scientific discoveries, stimulates new research and improves clinical care. It also provides a rich source of data that biopharmaceutical companies can use to augment their existing research and marketing activities.

There are two approaches biopharmaceutical companies can take to share data:

A. Online portals Data is made available through a portal that offers tools to do analysis but does not allow for data to be downloaded. Data may be open access or can have tighter controls that require users to register and agree to terms of use before being given access to the data.

- Open Trials (opentrials.net): An open-access, online database aggregating information that pertains to clinical trials from a variety of sources including regulatory documents, academic journals and other registers

- ClinicalStudyDataRequest.com (CSDR): An industry-created, controlled-access portal providing a centralized location for researchers to access anonymized patient-level data and supporting documents of clinical studies made available from various study sponsors.

B. Microdata Release In this case, patient-level data (i.e. IPD) from a clinical trial is shared with a data recipient as a database or flat file that they can download. Like online portals, a microdata release may be made publicly available (open access) or made available through controlled access requiring registration and agreement to terms of use.

- Project Data Sphere: A free digital library that provides controlled-access for researchers to

a single place containing academic and industry Phase III cancer clinical trials. The site represents more than 27,000 patient lives across a broad array of cancer tumor areas.

Different data sharing mechanisms present different risk profiles that require different levels of de-identification. The sensitivity of the data, the controls in place to protect it and the people who can access it are all factors that affect the data's context. The next section looks further at data's context and its impact on disclosing information.

Challenges and Concerns in Sharing Clinical Trial Data

Health data is highly sensitive, containing many personal details about patients that they would not wish to be shared. In making IPD or the CSR available, it is imperative that the data be anonymized to remove personal details so that it is highly improbable that an individual can be correctly identified from their information. When it comes to clinical trial data there are two issues that pose concerns for protecting patient anonymity: 1) the size of the dataset and 2) the context in which it is shared.

While many Phase III trials include thousands of participants from multiple sites worldwide, a clinical study conducted on a new treatment for a rare or complex condition may have fewer than 100 participants. Early trials (e.g. Phase I) can involve as little as a dozen individuals. Unfortunately, when we are dealing with small numbers of participants the risk of re-identification increases. Even if data is de-identified by redacting the direct identifiers like name or health card number, there is a greater chance of finding a person who has a unique combination of indirect identifiers (e.g. age, gender or race) within the dataset[3]. These unique cases are easier to re-identify.

In addition to the size of the dataset, the context in which data is shared also influences risk. How and where data is shared impacts its vulnerability to an attack. Data that will be made available to the general public, or data that can be obtained by anyone without added security or access controls, requires extensive de-identification. In this case, accessing the data does not require a person to register with the data provider, sign a data sharing agreement or undergo a security check. All of these techniques could be used to help protect the data by requiring knowledge of those who wish to use it and by making clear the repercussions an individual can face if they are responsible for a privacy breach. When these protections are lacking, shared data is more susceptible to a re-identification attack. Anyone with the inclination to launch a re-identification attack, provided he or she has the proper resources and know-how, can access the data for this purpose. Even data that is only available through controlled access mechanisms must be de-identified. The amount of de-identification applied to the dataset in this case, however, need not be as extensive since the risks to privacy are mitigated. Thus, context plays a significant role in assessing data to determine how much the specificity needs to be reduced. Greater specificity correlates with higher data quality.

By adopting de-identification processes that are in line with accepted best practices, study sponsors are protected as well. The sharing of IPD, even if it had been anonymized, may cause concern among some trial participants, particularly if they have not provided their express consent for it to be shared. As more and more

clinical trial data is shared, the likelihood of a complaint from a participant grows. A complaint filed with a regulator, like HHS, can trigger an investigation into a company's data sharing practices. Even if it is found that no privacy breach has occurred, the investigation could result in penalties or fines being levied against the company if it is found that data was not properly de-identified. Thus, even the lack of harm is no defense if a company is found to be lax in its data protection practices.

Knowing which organizations provide sound guidance for the responsible sharing of clinical trial data is critical as transparency initiatives and demand for secondary uses of data continue to increase. The next section reviews the current guidelines specific to disclosing clinical trial data.

Data Sharing Guidelines for Clinical Trials

Protecting the privacy of clinical trial participants is a necessary part of any plan to share study data. Recognition of this core principle is why there have been a number of guidelines developed by industry groups as well as independent organizations on the responsible sharing of clinical trial data.

In August 2014, the biopharmaceutical industry group, TransCelerate BioPharma, published their report, Clinical Study Reports Approach to Protection of Personal Data. The document outlines an approach that companies can use to protect patient privacy, specifically when sharing CSRs. The TransCelerate document advocates the use of data masking. It specifies that all personally identifiable information within a CSR must be de-identified by removal or redaction.

However, the problems with masking are two-fold; not only can masking overlook data that is identifying, it will also unnecessarily redact data that can be valuable in analysis, particularly for unstructured data[4]. This creates a concern for the reproducibility of studies and any additional analysis done since portions of the CSR contain unstructured data, including the description of serious adverse events in participants. Fortunately, there are other guidelines that promote a more robust approach to de-identification.

The Pharmaceutical Users Software Exchange (PhUSE) is a European-based, not-for-profit organization whose members work as biostatisticians, statistical programmers and data managers. Their standard for the de-identification of clinical trial data combines elements of the two de-identification methods cited in the US HIPAA legislation: Safe Harbor and Expert Determination. One of the benefits of the PhUSE guideline is that its approach is specific to de-identifying standard files (CDISC files) that are produced at the end of a clinical trial. As such, it indicates which variables should be classed as direct identifiers or which ones classed as quasi-identifiers, an important step in performing de-identification. Using a two-pass process, the direct identifiers are removed according to Safe Harbor recommendations. The PhUSE standard then calls for a residual risk analysis to be performed. If the re-identification risk is too high, a second pass is done using Expert Determination methodology.

The IOM issued their report Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk in January 2015. Developed with public and private-sector input, it provides guiding principles and a practical framework for the responsible sharing of

clinical trial data. A major focus of the document is the optimal timing to share different types of data from a clinical trial. However, it also notes that risks to privacy and security of data can be mitigated through the use of de-identification and data sharing agreements. It specifically notes the use of a risk-based approach, like HIPAA's Expert Determination method, for de-identification.

The last section discusses the implications of using the TransCelerate approach versus a risk-based approach as recommended by PhUSE and IOM and their impact on the meaningfulness of data sharing.

Achieving Meaningful Transparency

The quality of the resulting de-identified data is an important consideration when deciding how to anonymize it. Researchers and analysts require high-quality, granular data in order for their analyses to be accurate and meaningful. De-identification based on masking can greatly diminish data quality and undermine the purpose of sharing data. Companies that are committed to clinical trial transparency should consider following guidelines for anonymization that make use of a risk-based approach to de-identification, like the PhUSE standard and the recent anonymization guidelines released for EMA Policy 0070.

To date, however, there has been a lot of support for the TransCelerate approach from the biopharmaceutical industry. Not only does it provide a relatively simple set of rules for anonymizing CSRs, it also assuages industry concerns that greater transparency will lead to commercial confidential information (CCI) being revealed in reports. Since TransCelerate's method broadly removes and redacts content it is felt that it will effectively protect CCI, along with patient data.

The TransCelerate approach is based on Safe Harbor, a de-identification methodology documented under the HIPAA Privacy Rule. While Safe Harbor is applicable within the U.S. since it is part of U.S. legislation, it is not a globally-accepted standard for de-identification. Since biopharmaceutical companies operate internationally, they need to consider that their data anonymization is done in a way that is acceptable to regulators around the world. This makes TransCelerate less than ideal for initiatives outside of the U.S., including the new EMA policy.

Beyond jurisdictional concerns, the excessive redaction used in TransCelerate also poses a problem. The approach not only redacts direct identifiers like name and subject ID number but, goes beyond what is required by Safe Harbor to redact any patient-level demographic or socioeconomic information like sex, weight, height, race, ethnicity and occupation. And, while Safe Harbor requires all dates be generalized to year, TransCelerate redacts entirely all dates relating to an individual patient. This means that any information regarding a participant's birthdate, date of admission or date of intervention would be removed. Figure 1 (on the next page) shows sample data from a CSR with only the direct identifiers redacted. Figure 2, on the right, is the same sample redacted according to the TransCelerate guidelines. It is little more than a collection of black boxes. Add to this obstructed view of the data the removal of any full patient narratives and it is easy to see that the level of redaction would cripple the usefulness of the anonymized document.

CRTN/Pt. No.	Age yr	Sex	Weight kg	Height cm	Race	Treatment Start End
[redacted]3030	34	F	65	[redacted]	CAUCASIAN	18JAN2001 22JAN2001
[redacted]3032	13	F	81	[redacted]	CAUCASIAN	22JAN2001 26JAN2001
[redacted]3033	37	F	76	[redacted]	CAUCASIAN	13FEB2001 17FEB2001
[redacted]3035	43	M	98	[redacted]	CAUCASIAN	30JAN2001 04FEB2001
[redacted]3045	20	M	57	[redacted]	CAUCASIAN	30JAN2001 04FEB2001
[redacted]3046	20	F	55	[redacted]	CAUCASIAN	23FEB2001 27FEB2001
[redacted]3048	21	M	75	[redacted]	CAUCASIAN	10FEB2001 14FEB2001
[redacted]3049	24	F	45	[redacted]	CAUCASIAN	02FEB2001 06FEB2001
[redacted]3050	47	M	90	[redacted]	CAUCASIAN	08FEB2001 12FEB2001
[redacted]3271	21	M	88	[redacted]	CAUCASIAN	22FEB2001 26FEB2001
[redacted]3272	34	M	113	[redacted]	CAUCASIAN	07MAR2001 11MAR2001
[redacted]3274	25	F	64	[redacted]	CAUCASIAN	08MAR2001 12MAR2001
[redacted]3060	29	M	59	[redacted]	CAUCASIAN	20FEB2001 24FEB2001
[redacted]3062	43	F	60	[redacted]	CAUCASIAN	16FEB2001 20FEB2001
[redacted]3090	51	M	95	[redacted]	CAUCASIAN	14FEB2001 18FEB2001
[redacted]3092	38	F	75	[redacted]	CAUCASIAN	08FEB2001 13FEB2001
[redacted]3101	28	F	62	[redacted]	CAUCASIAN	19FEB2001 24FEB2001
[redacted]3104	26	F	55	[redacted]	CAUCASIAN	09FEB2001 13FEB2001
[redacted]3105	28	F	70	[redacted]	CAUCASIAN	08FEB2001 13FEB2001

Figure 1: A CSR sample with direct identifiers redacted

CRTN/Pt. No.	Age yr	Sex	Weight kg	Height cm	Race	Treatment Start End
[redacted]3030	[redacted]		[redacted]	[redacted]	[redacted]	[redacted]
[redacted]3032						
[redacted]3033						
[redacted]3035						
[redacted]3045						
[redacted]3046						
[redacted]3048						
[redacted]3049						
[redacted]3050						
[redacted]3271						
[redacted]3272						
[redacted]3274						
[redacted]3060						
[redacted]3062						
[redacted]3090						
[redacted]3092						
[redacted]3101						
[redacted]3104						
[redacted]3105						

Figure 2: The same CSR sample masked according to the TransCelerate Guidelines

If CSRs are censored to the point that they are unreadable, then there is limited value in publishing them as a means of bolstering transparency. More information could be retained in this situation by using de-identification techniques like aggregation and date shifting, tools that are available with risk-based de-identification. Ages could be aggregated so that the participant's age is shown as 20-29, 30-39, etc. Similarly, dates could be shifted to preserve information about the treatment duration without revealing the actual treatment dates. Both of these approaches remove some of the data's specificity, making it more difficult to positively re-identify a specific individual[5]. It has been shown, however, that data de-identified in this way is sufficient to reproduce original study results.

The EMA's stance has been one of maximizing data utility. Their Anonymization Guidance for CSRs states the importance of maximizing the amount of scientifically useful information in clinical reports. This guidance also stipulates that the methodology include a way of measuring re-identification risk and having a repeatable process to follow. Thus, in addition to providing an anonymized CSR for release, companies must also provide an anonymization report that describes the methods used and their impact on data quality, information that can readily be produced if a well-designed de-identification process is in place.

Patient anonymity can be achieved while providing high-quality data and protecting CCI.

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

By following standards that take a risk-based approach to de-identification, clinical study sponsors can provide anonymized documents that meet all aspects of the EMA's new requirements for clinical trial transparency and that comply with privacy regulations in various jurisdictions.

For a complete explanation of how a clinical trial dataset can be de-identified according to the PhUSE standard, see the Privacy Analytics' webinar, A Case Study of De-identifying a Clinical Trial Dataset.

Conclusion

The EMA's policy 0070 is an unprecedented attempt to enhance the public's trust and confidence in the drug approvals process. However, requiring publication of a study's anonymized CSR without specifying how the data should be de-identified could compromise the meaningfulness of this initiative. Publishing a regulatory document does not make sense if it has been so excessively redacted that the facts contained in it are compromised. If CSRs are censored to the point that the readability and reproducibility of the trial is jeopardized, it could nullify progress towards clinical trial transparency.

While it is still early days for the new EMA requirements, forward-thinking biopharmaceutical companies are seeing the opportunity to position themselves as leaders in sharing clinical trial data. By using robust anonymization processes based on risk-based de-identification, these companies can provide high-quality CSRs that meet the intent of the EMA policy and deliver true transparency. With a methodology that is effective for both CSRs and IPD, companies that embrace risk-based de-identification will be positioned to scale their future activities as the need dictates.

Sources:

1. U.S. National Institutes of Health. Trends, Charts and Maps. ClinicalTrials.gov.
2. Clinical Study Data Request. ClinicalStudyDataRequest.com.
3. For a more detailed explanation, see equivalence classes in the white paper De-identification 301: Three Adversaries Who Could Attack Your Data
4. For more on data masking see the white paper, Avoid the Blur of Data Masking
5. For more on de-identification techniques, see white paper De-identification 201: Fundamental of Data De-identification