CRDSA

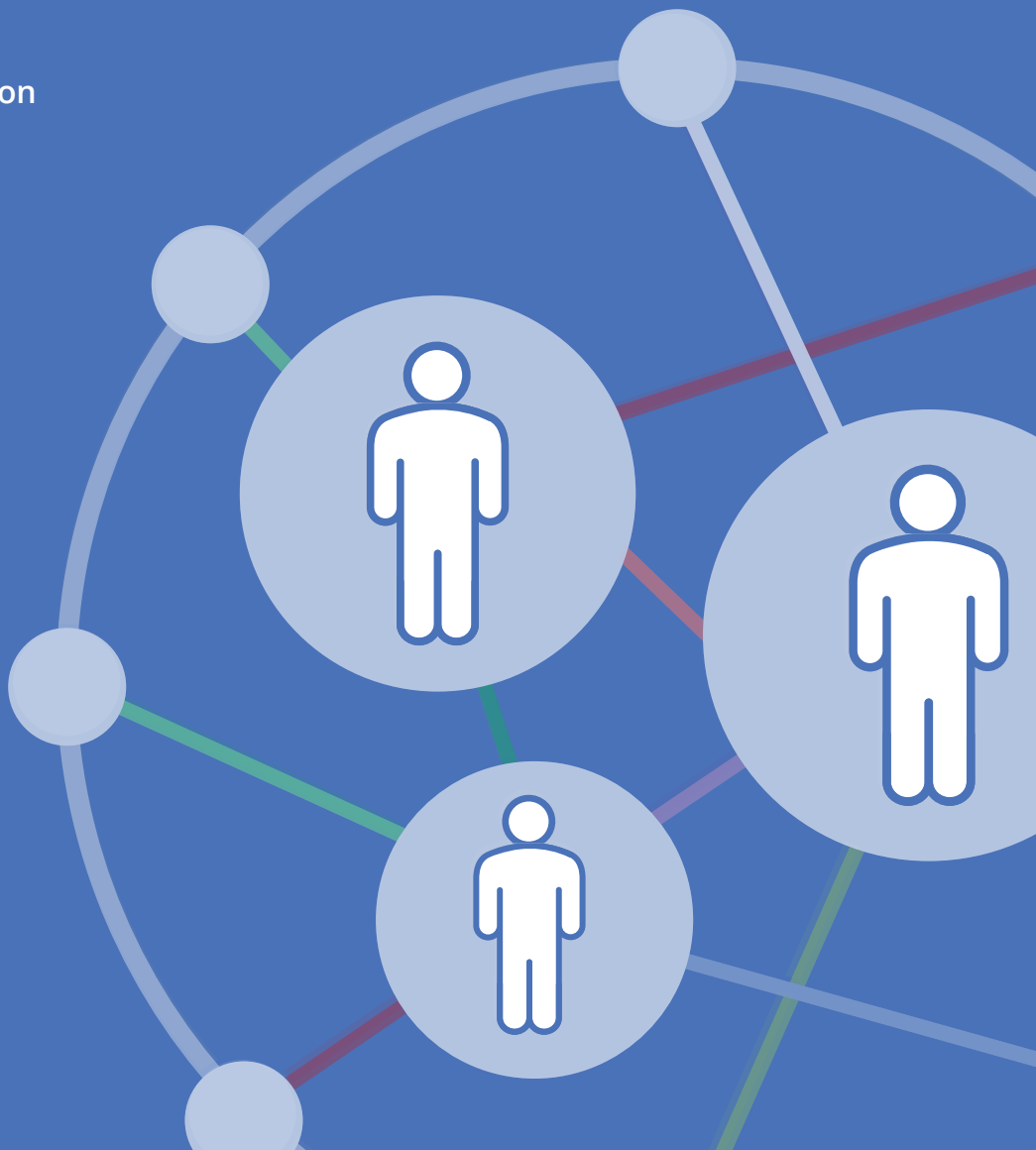
Clinical Research Data Sharing Alliance

A Review of BioPharma Sponsor Data Sharing Policies and Protection Methodologies

Version: 1.0

Work Group: Data Protection

Date: 12 September 2022



A Review of BioPharma Sponsor Data Sharing Policies and Protection Methodologies

Liz Roberts (UCB), Luk Arbuckle (Privacy Analytics), Andrei Belcin (Privacy Analytics),
Chad Burris (Takeda), Cathal Gallagher (D-Wise), Aaron Mann (Clinical Research Data Sharing Alliance)
on behalf of the CRDSA Data Protection Work Group

Check for Updates »

Abstract

This whitepaper examines clinical trial data contribution policies and the data protection methodologies applied to protect patient privacy. Information published by 29 biopharma sponsors was collected across three data-sharing platforms, collated by sponsor size. Results showed that large sponsor contribution policies can provide helpful benchmarks for medium and smaller sponsors. However, understanding the data protection methodologies applied by data contributors required the development of an interpretation rubric. This lack of clarity in the critical data transformation area can present a significant barrier for researchers and hinder their understanding of which studies will be useful in their research. Findings of the paper include highlighting an opportunity for sponsors to both aid researchers and reduce unnecessary data preparation work by sharing information with more consistency and clarity.

Acknowledgments

The Clinical Research Data Sharing Alliance would like to thank the following people for their support:

CRDSA Data Protection Work Group:

Louise Dexter (AstraZeneca), Stephen Doogan (Real Life Sciences), Khaled El Emam (Replica Analytics), Ashley Hopkins (Flinders University), Michael Sorich (Flinders University), Priya Pavithran (GSK)

Reviewers:

Jessica Lim (GSK), Peter Mesenbrink (Novartis), Frank Rockhold (Duke Clinical Research Institute), Julie Wood (Vivli)

Introduction

Sharing patient data generated from clinical trials is fundamental to the advancement of science and the improvement of public health. The use cases for shared data are well-established and include novel clinical trial design and enrichment strategies, predictive preclinical and clinical models, clinical trial simulation tools, biomarkers, clinical outcomes assessments, and more.[1] Shared clinical research data also has the power to transform the trial process itself, improving the patient experience and delivering life-saving and life-changing therapies faster and at less cost to society.

However, when sponsors put data-sharing into practice, there can be high variability in both contribution volume and utility to end-users. It's important to recognize the challenges faced by data contributors. Ensuring contributions maximize research utility must balance with the equally important need to responsibly protect patient privacy. This balancing act results in a wide spectrum of contribution approaches, some of which may compromise data utility to the point where scientifically interpretable analyses become increasingly challenging.

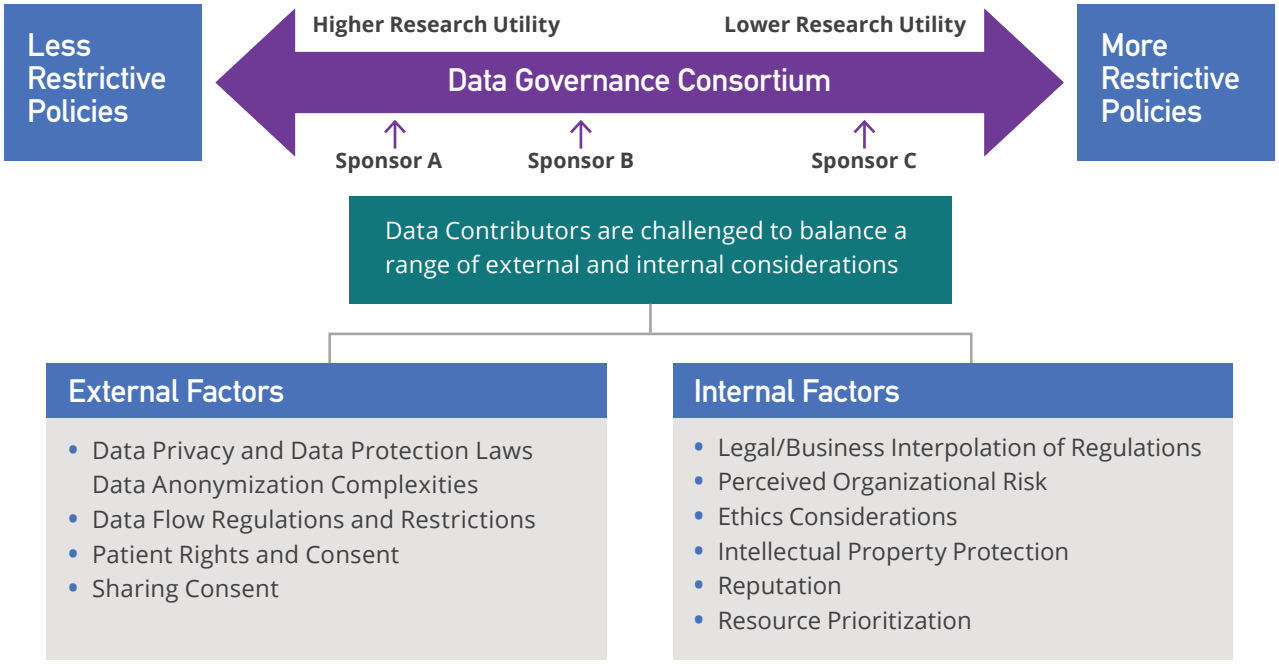


Figure 1: Data Governance Continuum

This whitepaper focuses on elements impacting end-user research utility, including what data is shared and how data will be transformed to protect patient privacy. Through this review, the authors seek to create meaningful benchmarks to help guide data contributor policy development.

Since 2015, Bioethics International has published the Good Pharma Scorecard[2]. The scorecard, published bi-annually, assesses data transparency and, in 2021, added an evaluation of clinical trial data sharing to the report. Biopharma sponsor data sharing practices and policies are evaluated across five criteria. The scorecard's first data sharing criterion is "whether they have a public policy committing to sharing analysis-ready datasets and clinical study reports (CSRs) for applicable studies."

This criterion begins to address the important question of "what is shared" by data contributors. The datasets and supporting documentation contributed are important determinants of utility to end-users. They provide researchers with the critical context and underlying information needed to address scientific inquiries. The authors' review builds on the Good Pharma Scorecard criterion by expanding the datasets and documentation evaluated and the addition of an analysis of applied data protection methodologies.

Methodology

Source Data

The source data for this review was collected from publicly available information published by trial sponsors across three data sharing platforms: Vivli, Clinical Study Data Request (CSDR), and the Yale Open Data Access (YODA) Project [Appendix A]. The information sponsors publish on these platforms can include informational statements, data points (e.g., lists of supplied documents), and linked in-depth policy and/or process documents.

Of the sponsors listed on these platforms, the 29 biopharma sponsors provided information sufficient for this analysis, while most academic sponsors did not provide detailed information. To provide meaningful comparisons, this review focuses on the information provided by the biopharma sponsors.

The authors are aware that some of the sponsor information may be out of date. However, it was determined that our analysis would rely on publicly available information published on the data sharing platforms as of our review period (April 15-May 30, 2022). Critically, the published sponsor information is what's available to researchers. We encourage sponsors to review available information to ensure it is consistent with their current policies and practices.

Data Protection Assessment Methodology

In addition to synthesizing the information supplied, the team developed a consistent methodology to categorize data protection approaches. The objective of applying this methodology was to determine whether a sponsor uses a risk-based data protection approach (i.e., objectively supported through measures of identifiability, or the risk of re-identification, and a reasonableness standard). A risk-based approach to data protection is generally thought to improve the end-user utility of data contributions[3].

In our analysis, we encountered divergences in what is considered risk-based anonymization, similarly shown in Clinical Trials[4], which conducted a literature review on recommended clinical trials anonymization (or de-identification) approaches to determine what was consistent across the recommendations made; as a result, there is no single standardized set of recommendations related to anonymizing clinical trial datasets for sharing. The review also suggested that the researchers surveyed considered anonymization techniques, including risk-based approaches, insufficient to protect patient privacy unless they factor in trusted methods for controlled access to the anonymized data.

The role of controlled access on the degree of anonymization, especially the effect on risk-based, is depicted below by the SAFE data standard[5], published after the literature review with the intent of promoting standardization and efficiency in the sharing of clinical trial data:

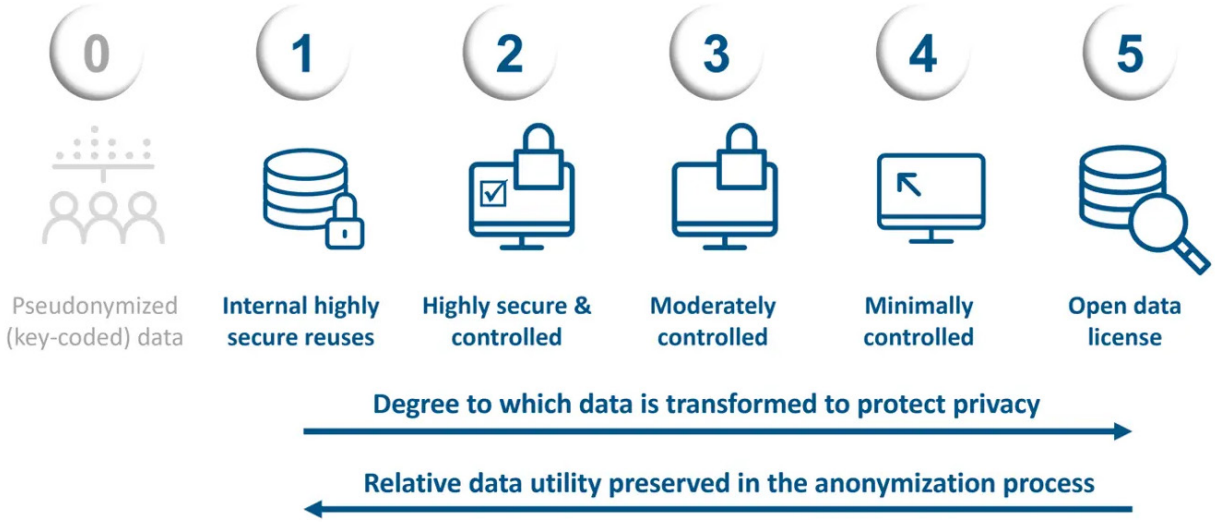


Figure 2: *SAFE data rating by Bamford et al. (Applied Clinical Trials, 2022)*: Description of the SAFE Data Standard rating system, where data rated from 1 to 5 has been anonymized to reflect the context of data disclosure, with an increasing degree of data transformation and associated impact on data utility. "Internal" refers to the trial sponsor reusing its data.

The data rating applied informs the degree of anonymization required to ensure a re-identification probability below a threshold where often platforms, including the platforms in this review (Vivli, the YODA Project, and CSDR), may rate at 2 (depending on implementation), internal use within the sponsor organization is often rated at 1, whereas case-by-case partnerships with individual organizations (not sharing platforms) often score a 3 on the scale above. To classify approaches with confidence, the authors scaled expectations of the sponsors' published standards to reflect a rating of 2 (i.e., highly secure and controlled) for the platforms included in this review.

From the publicly available anonymization standards by biopharma sponsors on Vivli, CSDR, and the YODA Project sharing platforms, we can observe three criteria for evaluating a standard to classify it as "risk-based" (in decreasing order of confidence):

1. The extent an approach to anonymization is modified based on specific factors influencing risk, e.g., study population, disease prevalence, data sensitivity, system controls, context risk, various re-identification attack scenarios, and adversary profiles
2. Explicit claim by the sponsor that their anonymization approach is "risk-based"
3. A mention of a "risk assessment" in addition to a specified rule-set

These criteria were determined from both domain knowledge and the tendencies in specifications among the sponsor's published standards scaled to the expected access controls of the sharing platforms. The review used the following categorizations in assessing the data protection methodology used by a sponsor:

- **Criterion 1.** is classified as using a risk-based anonymization approach, and we can also infer the use of a quantitative approach
- **Criterion 2.** is classified as using a risk-based anonymization approach
- **Criterion 3.** is possibly a risk-based anonymization approach
- **Else,** if a sponsor uses a rule set minus the risk assessment component, then we infer a rules-based approach used

As our research centered around public domain documentation from sponsors, our ability to assess the details was limited to what was explicitly available. The authors did not bring forward any knowledge of current practices by sponsors. Based on confidential knowledge among the group of authors, however, it is evident that many sponsors have outdated documentation describing their anonymization standards. Exacerbating the situation, there are still a number of sponsors not providing any documentation on their anonymization approach.

Few sponsors stated their assessment of re-identification (disclosure) risk involved various factors, *Criterion 1, of which we infer to include some form of quantitative measurement* of risk and in turn informs the amount of transformation on the trial data to achieve an expected utility by the data recipients (researcher, regulators). Some sponsors provide a short (2 -3 pages) document which makes claims about their approach without sharing the set of transformations considered and the basis for choosing to enact them on the clinical trial data at a high level; as a result, the authors of this assessment included *Criterion 2, a claim of using a risk-based approach*. Similarly, many sponsors generally use the phrase "risk assessment" preceding the outline of their rules-based anonymization and without any link of what the assessment entails and how it affects the degree to which rules are applied to trial data. This leads us to form *Criterion 3 to give sponsors the benefit of the doubt that they may, in fact, be using a risk-based approach*. Lastly, the else clause *captures all other approaches in which* there is no evidence, no claim, and no risk assessment to even suggested that a risk-based approach is used.

Collation and Interpretation

There is inherent variability in the underlying information used in this review. As noted above, some information may not be current. It was also noted that sponsors might use terms differently and/or interchangeably (e.g., anonymization and de-identification). Further, our analysis, particularly regarding the applied data protection methodology, is based on a best-efforts interpretation of the supplied information. For these reasons, the authors have chosen to present this review at the aggregate level, tiered by sponsor size.

The direct and indirect (interpreted) information was collated into three tiers based on sponsor size. Tiers were determined using the total sponsor employee count as follows:

Tier 1	Tier 2	Tier 3
Over 25,000 employees	**5,000 to 24,999 employees**	**4,999 employees or fewer**

The breakpoints between tiers were distinct, and no sponsors were within 15% of an adjacent tier. Of the 29 sponsors included in this review, 12 fall in Tier 1, 11 in Tier 2, and 6 in tier 3. The list of reviewed sponsors is provided in Appendix A.

Results

Datasets and Documentation

Tier 1: 25k + Tier 2: 5 to 24.99k Tier 3: Under 5k	Tier 1 (n=12)		Tier 2 (n=11)		Tier 3 (n=6)	
Datasets and Documentation Shared						
Raw	12	100%	8	73%	5	83%
Analysis	11	92%	9	82%	4	67%
Protocol	12	100%	8	73%	5	83%
Annotated CRF	12	100%	7	64%	4	67%
Reporting and Analysis Plan / SAP	12	100%	8	73%	4	67%
CSR	11	92%	9	82%	2	33%
Dataset Specifications	9	75%	7	64%	3	50%

Table 1: Datasets and Documentation
(Reference Appendix B for additional resources including information on dataset and documentation types.)

Larger Tier 1 sponsors were more likely to share datasets and documentation than smaller Tier 2 and 3 sponsors. This increased sharing extends across all data and documentation types, with the greatest differences in annotated CRF sharing, with all Tier 1 sponsors sharing compared to 2/3 or less of Tier 2 and 3 sponsors. The smallest differences between tiers are for dataset specifications, which were the least likely component to be shared by Tier 1 and 2 sponsors. All study documentation components were shared by at least half of the sponsors in each tier except for CSRs, which were shared by only 1/3 of Tier 3 sponsors.

Larger sponsors' increased propensity to share more study documentation most likely results from increased experience and internal expertise with the European Union General Data Protection Regulation ("GDPR") and other privacy regulations which can be applied to clinical trial data sharing. Larger sponsors also command more resources and are more likely to have the ability to direct

resources toward data sharing. At the same time, due to their bolus of completed trials, larger sponsors are likely to field a correspondingly high number of data requests and, as a result, may have more experience in processing these requests. In contrast, smaller sponsors may have a rare disease focus which poses additional challenges for document redaction given that rare disease data may make it harder to protect patient privacy.

While all Tier 1 sponsors shared study protocols, only 73% of Tier 2 and 83% of Tier 3 sponsors shared protocols. The lack of protocol sharing by Tier 2 and 3 sponsors may be partially explained by concerns over the amount of redaction necessary resulting from the more significant amount of redaction needed when submitting data to clinicaltrials.gov[6] compared to a lighter redaction burden for controlled data sharing platforms.

Further to the table above, it was noted that the number of sponsors sharing ***raw*** data sets did not equal the number sharing ***analysis*** level data sets.

Tier 1: 25k + Tier 2: 5 to 24.99k Tier 3: Under 5k	Tier 1 (n=12)		Tier 2 (n=11)		Tier 3 (n=6)	
Those sharing Raw but not Analysis data	2	17%	2	18%	1	17%
Those sharing Analysis but not Raw data	0	0%	2	18%	0	0%

Table 2: Raw/Analysis Comparison

It is more common for some sponsors to share raw data and not the associated analysis than vice versa. This may stem from limited resources available for data anonymization with a resulting choice to only provide either raw or analysis data instead of sharing both. Those sharing only raw data may believe the raw data is most useful for secondary analysis (e.g., to facilitate pooling across data sets), while those sharing only analysis data may hold the opposite belief, prioritizing access to analysis-ready data sets. Ideally, both raw and analysis data should be shared when possible so that researchers have analysis-ready data with complex calculated endpoints available while also having the ability to research how these were derived so they can determine the applicability to researchers' secondary analyses.

There is an opportunity to support smaller sponsors who may lack the resources and expertise found at larger sponsors by helping them to adopt policies that facilitate clinical trial data sharing and lead to improvements in the number of smaller sponsors sharing full supporting documentation.

Data Protection Methodology

Techniques	Tier 1^ (n=12)		Tier 2^ (n=11)		Tier 3^ (n=6)	
Risk Based approach Distinctly described or highly likely	5	42%	3	27%	1	17%
Rules Based Approach	5	42%	5	45%	2	33%
HIPAA/Other Approach	2	17%	3	27%	3	50%
Anonymization Terminology						
Uses "Anonymized"	10	83%	10	92%	3	50%
Uses "De-Identified"	1	8.5%	0	0%	2	33%
Uses both interchangeably	1	8.5%	0	0%	0	0%
Cannot determine (missing blurb or documentation)			1	8%	1	17%
Additional Documentation (Linked Documents)						
Sponsors that provided additional methodology detail	8	67%	6	55%	3	50%
Subset where detail outlines a similar data protection methodology*	5	42%	4	36%	3	50%

Table 3: Data Protection Methodology

^ Tiers are defined by number of employees. Tier 1 25k+, Tier 2 5k – 25k, Tier 3 <5k

* Where additional detailed documentation was provided, Appendix C summarizes the general methodology used solely as the rules-based approach or in conjunction with a risk assessment.

To expand on the sponsors' degree of risk-based anonymization, the table below highlights the distribution of the assessment criteria across the three tiers.

Techniques	Tier 1^ (n=12)		Tier 2^ (n=11)		Tier 3^ (n=6)	
Risk Based approach	5/12	42%	3/11	27%	1/6	17%
Criterion Breakdown						
Criterion 1 - The extent an approach to anonymization is modified based on specific factors: study population, disease prevalence, data sensitivity, system controls, context risk, various re-id attack scenarios, and adversary profiles	1	20%	0	0%	1	100%
Criterion 2 - Reference to a "risk assessment" in addition to a specified rule-set	2	40%	2	67%	0	0%
Criterion 3 - Stated by the sponsor that their anonymization approach is "risk-based"	2	40%	1	33%	0	0%

Table 4: Sponsors by Data Protection Criterion Used

The most descriptive use of risk-based anonymization is detailed in criterion 1. The research illustrates that this is an outlier amongst published standards from all Tiers. Criteria 2 and 3 are similarly preferred by sponsors when describing their risk-based methodology, although they each contain ambiguities that do not definitively declare that the anonymization approach is quantitatively risk-based. The authors recommend sponsors outline the basis on which their approach to anonymization can be labeled as risk-based; as sponsors relying on Criterion 3 offer no basis, and while sponsors relying on Criterion 2 make a vague link between an undefined "risk assessment" and what (if any) effect it has on determining the rules applied on the data to "minimize risk of disclosure."

Additional Dimensions

Sponsors also provide information about their general data sharing policies regarding what studies will be shared, when studies become eligible for secondary use data sharing, and exceptions. As above, the authors collated the information by sponsor size tier. While large sponsors tended to give somewhat more context to the information presented, we did not find significant policy differences between tiers. Therefore, these results are presented as applicable across tiers.

What studies are shared?

Most sponsors share Phase 2-4 interventional studies (some provided no information). Six of the 29 sponsors (21%) also indicate they share Phase 1 studies; however, it should be noted that most Phase 1 studies would fall under a study exception (see below) due to the limited number of patients. The studies generally in sharing scope have been through a regulatory approval process (typically approval in one or more of US, EU, and Japan) or if the development product is terminated. We did observe that some smaller sponsors (Tier 3) require marketing approval before sharing.

When are studies shared?

Of the total 29 sponsors surveyed, 22 provided information on timelines for study sharing eligibility. The majority of sponsors providing information (12 of 22) will share trials 18 months after study completion or 6 months after publication. This is consistent with the Good Pharma Scorecard's benchmark criterion: "whether the policy commits to making data available by 6 months after approval by the FDA or European Medicines Agency or 18 months after a trial's completion date, whichever was later." 6 of the 22 (27%) sponsors indicate a more liberal (i.e., faster time to share) policy than the scorecard criterion, while 4 (18%) indicate a stricter sharing policy.

Study Exceptions

Study exceptions were generally consistent across sponsors. The most common reason study data cannot be shared is an inability to achieve an anonymization threshold that adequately protects patient privacy. This can be because of a small patient population (typically under 50), rare disease indication, or geographic considerations (e.g., a single-center trial). Studies that have been co-developed or are co-owned are also frequently cited as generally out of scope.

Data Exceptions

Data exceptions were also generally consistent across sponsors, including imaging (x-rays, MRI scans), genetic data, exploratory biomarkers, and non-English documents. It should be noted that generally accepted best practice data protection methodology for some data types, including images and genetic data, is still being developed.

Discussion and Conclusions

The aim of this whitepaper was to focus on elements impacting end-user research utility, including what data was in-scope for sharing and how that data would be transformed to protect patient privacy.

Regarding scope, this was reasonably consistent across data contributors. Datasets and associated documents from Phase 2-4 interventional clinical trials (sometimes also Phase 1 trials) were generally in scope for sharing. This sharing occurred after data lifecycle milestones were met, including a certain number of months after the end of the trial, the primary publication date, and/or regulatory approval. Other forms of data, such as images, genetic data, and non-English documents, were normally not in the scope of sharing.

It was much less clear to understand the types of protections applied to the data prior to sharing and how the risk of re-identification was mitigated. These are important facts for researchers so that it is clear whether the data will be available in a form that allows analyses to proceed as planned, i.e., for the shared data to have clinical utility.

Through the creation of this whitepaper, the authors seek to identify meaningful benchmarks to guide Data Contributor policy development, including the required clarity on such policy.

Recommendations

1. Clarity of Available Information

- The terms 'de-identified' and 'anonymized' are sometimes used synonymously, although they have different meanings in different regions. Alignment of terminology and their definitions would provide much-needed clarity across the global data sharing ecosystem.
- Data Contributors could provide more information upfront explaining their approach to data protection. This is especially important when researchers plan to combine data from multiple contributors and need clarity to understand if such data can be pooled. For example:
    - Will any data, such as rare adverse events, need to be removed prior to sharing? If this high-level fact is transparent, then researchers whose analysis requires such data would be able to identify such trials as potentially not being able to provide sufficient data utility.
    - If the research requires granular demographic data, is it clear how these variables will be transformed (or whether even present) in the datasets that are to be shared?
    - If the research is in an area where seasonality is an important consideration, how will this have been addressed in the datasets that are to be shared?
- Some sponsors describe the extent to which their approach to data protection is modified based on specific factors, including study population, disease prevalence, data sensitivity, system controls, context risk, various re-identification attack scenarios, and adversary profiles. This gold-standard approach provides much more clarity to researchers and is also aligned with the preferred approach by regulators as part of mandatory document publication policies.

Sponsors: Recommended Anonymization Methodology Detail

Overall, the authors would recommend that sponsors across tiers provide the following details as baseline components in their anonymization standards on clinical trial data sharing platforms:

- Specificity on the risk assessment; application of quantitative or qualitative methodology; and the relevant factors considered in the assessment
- The way in which the risk assessment informs the transformations of the clinical trial data to anonymize it
- References to, or in support of, the methodology used to anonymize and produce useful data for end-users

As a result, the recipients of the data can both form their own judgment on how confident they are in the data meeting patient privacy requirements and build an expectation of the utility of the data. As standards emerge and validation of data integrity becomes more common, end-users will be more likely to trust the data and insights derived, improving the coordination of health research and increasing public trust.

2. Benchmarking

- The responsible sharing of patient-level data is not intended to be a box-ticking exercise conveying compliance with various regulations and guidelines. To support meaningful data sharing, Data Contributors must ensure processes are followed to protect individuals' data privacy but also retain as much data utility as possible in the resultant datasets and documents that are to be shared. If there is insufficient transparency (data are not 'FAIR'[7]), then the potential secondary benefit is lost. If requested data sources do not contain sufficient data utility, then time and money are lost by both the research team and the Data Contributor. It is also possible that the omission of that data could have deleterious effects on meta-analyses, including drawing incorrect inferences based on incomplete or non-representative data sets.
- Benchmarking is an important tool for Data Contributors as it allows organizations to understand where they are relatively placed in the data sharing ecosystem. It can provide information related to return on investment, including:
  - How many (and which) clinical trials are requested (and how often);
  - whether the data in these trials can be used to generate new research as planned (e.g., Does it contain the required utility? Do research projects tend to complete or are they abandoned?); and
  - knowledge of the new scientific insights which could lead to new potential treatments.

There is an opportunity for the research community and the data contributor community (recognizing that many organizations are active in both fields) to come together and discuss pain points and opportunities. There are likely several quick wins that could be identified, and the scope to explain why they are not possible if process changes cannot be accommodated. By bringing the two communities together, there is the potential to create a paradigm shift within the data sharing ecosystem that more efficiently connects researchers with highly usable data and also benefits Data Contributors as they can be more confident that the shared data will actually be used. This is a win-win-win scenario – a win for researchers, a win for Data Contributors, and most importantly, a win for patients who rely on researchers to identify new scientific insights that ultimately could lead to new treatments.

About CRDSA

The Clinical Research Data Sharing Alliance (CRDSA) is a multi-stakeholder consortium serving the clinical data sharing ecosystem. CRDSA's mission is to accelerate the discovery and delivery of life-saving and life-changing therapies to patients by expanding the research value of the high-quality data collected through the clinical trial process.

References

1. Karpen, S.R., White, J.K., Mullin, A.P. et al. Effective Data Sharing as a Conduit for Advancing Medical Product Development. Ther Innov Regul Sci 55, 591–600 (2021). https://doi.org/10.1007/s43441-020-00255-8: 591, 599
2. Axson SA, Mello MM, Lincow D, et al. Clinical trial transparency and data sharing among biopharmaceutical companies and the role of company size, location and product type: a cross-sectional descriptive analysis. BMJ Open 2021;11:e053248. doi: 10.1136/bmjopen-2021-053248
3. Polonetsky J, Tene O, Finch K. Shades of gray: seeing the full spectrum of practical data de-identification. Santa Clara Law Review. 2016;56:593–629.
4. Rodriguez A, Tuck C, Dozier MF, et al. Current recommendations/practices for anonymising data from clinical trials in order to make it available for sharing: A scoping review. Clinical Trials. June 2022. doi:10.1177/17407745221087469
5. SAFE data rating. Reprinted from Applied Clinical Trials by Bamford, S., Lyons, S., Arbuckle, L., & Chetelat, P. (2022). Sharing Anonymized and Functionally Effective (SAFE) Data Standard for Safely Sharing Rich Clinical Trial Data. Applied Clinical Trials, 31(7/8). https://www.appliedclinicaltrialsonline.com/view/sharing-anonymized-and-functionally-effective-safe-data-standard-for-safely-sharing-rich-clinical-trial-data
6. U.S. National Library of Medicine. (2022) ClinicalTrials.gov. https://clinicaltrials.gov/
7. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Appendix A: List of Platforms and Sponsors

Platforms

- Clinical Study Data Request (CSDR) (https://www.clinicalstudydatarequest.com/)
- Vivli (https://vivli.org/)
- Yale Open Data Access (YODA) Project (https://yoda.yale.edu/)

Sponsors

- Alnylam Pharmaceuticals (https://www.alnylam.com/)
- Astellas (https://www.astellas.com/en/)
- AstraZeneca (https://www.astrazeneca.com/)
- Bayer (https://www.bayer.com/en/)
- Biogen (https://www.biogen.com/en_us/home.html)
- Boehringer Ingelheim (https://www.boehringer-ingelheim.com/)
- Bristol Myers Squibb (https://www.bms.com/)
- Daiichi-Sankyo (https://www.daiichisankyo.com/)
- Eisai (https://www.eisai.com/index.html)
- Lilly (https://www.lilly.com/)
- GSK (https://www.gsk.com/en-gb/)
- Grunenthal (https://www.grunenthal.com/)
- Johnson & Johnson (https://www.jnj.com/)
- Kyowa Kirin (https://www.kyowakirin.com/index.html)
- Lundbeck (https://www.lundbeck.com/us)
- Mitsubishi Tanabe Pharma (https://www.mt-pharma.co.jp/e/)
- Novartis (https://www.novartis.com/)
- Otsuka (https://www.otsuka-us.com/)
- Pfizer (https://www.pfizer.com/)
- Regeneron (https://www.regeneron.com/)
- Roche (https://www.roche.com/)
- Sanofi (https://www.sanofi.com/)
- Shionogi (https://www.shionogi.com/us/en/)
- SpecGx LLC, a subsidiary of Mallinckrodt Pharmaceuticals (https://www.mallinckrodt.com/)
- Sumitomo Pharma/Sunovion Pharmaceuticals, Inc. (https://www.sunovion.com/)
- Taiho Pharmaceutical (https://www.taihooncology.com/)
- Takeda (https://www.takeda.com/)
- Tempus (https://www.tempus.com/)
- UCB (https://www.ucb.com/)

Appendix B:

Additional Resources and References

Vivli:

Video: Why is the Data Anonymized/De-Identified Prior to Sharing?
https://youtu.be/PzX5cMCQ3XI

Video: What are the Supporting Documents Provided along with IPD?
https://youtu.be/dBE1hUDvWb8

Video: What is a Clinical Study Report (CSR)?
https://youtu.be/ozRJBMJOWBI

Video: What is an analysis-ready dataset (AD)?
https://youtu.be/Xm3_w_sEG28

Yale Open Data Access (YODA) Project:

Article: Sharing clinical trial data: lessons from the YODA Project
https://www.statnews.com/2019/11/18/data-sharing-clinical-trials-lessons-yoda-project/

PHUSE:

Whitepaper: Terminology Harmonisation in Data Sharing and Disclosure Deliverables Terms and Definitions V2
https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/WP065.pdf

TransCelerate BioPharma:

Toolkit: A Privacy Framework for Clinical Data Reuse: Secondary Data Use in the Pharmaceutical Industry
https://www.transceleratebiopharmainc.com/assets/interpretation-of-clinical-guidances-regulations-solutions/#gdpr-data-reuse

Appendix C:

Table 3: Data Protection Methodology

Where sponsors provided additional detailed documentation, the following summarizes the general methodology outlined:

1. Remove personally identifiable information from the dataset of the 18 identifiers (as defined by HIPAA US).
2. Recoding identifiers and research subjects' identification code numbers.
3. Removing free text verbatim terms.
4. Replacing date of birth by age (banded).
5. Replacing all original dates relating to a study subject using either: a) dummy date method or b) study day offset method.
6. Reviewing and Removing/Redacting Other PII: sites/labs, locations, investigators, imaging data (MRI, x-ray), etc
7. Quality control checks on the anonymization and packaging of the data/documents to be shared in separated locations from the original data.
8. Destroying link (key code) between de-id data and source trial data, storing anonymized data separately from the source data, and erasing remnants of processing the source data.