CLINICAL DATA SHARING: A PROPOSED METHODOLOGY TO ENABLE DATA PRIVACY WHILE IMPROVING SECONDARY USE

AUGUST 2023

FINAL VERSION

TransCelerate BIOPHARMA INC.
Accelerating Answers.

Contents

1.0 INTRODUCTION

As part of clinical research, pharmaceutical companies collect data related to the health and well-being of individual human beings. Sharing and combining clinical trial participant data for further research ("secondary use/research") has the potential to increase our scientific understanding, develop new medical treatments, and improve healthcare. There is a willingness among clinical trial participants to share their data for secondary research purposes when possible [1]; however, there are also privacy concerns that need to be addressed, requiring adequate safeguards to be in place prior to any data sharing activities.

To anonymize[i] personal data across multiple data sources, state-of-the-art measures[ii] must be considered. But what does this mean in practice? The goal of this paper is to propose a data privacy methodology for reuse of clinical trial data in research that preserves the privacy of research participants and maintains the scientific integrity and utility of the clinical data.

Most countries consider research participant data as personal data. Use of personal data for research must ensure the confidentiality of the research participants, protect their rights, and honor their trust in pharmaceutical companies when taking part in industry-sponsored studies.

What is required to protect privacy can be challenging to determine, especially because any data privacy approach must comply with an evolving legislative and regulatory landscape. Across the pharmaceutical industry, there is a general acceptance that data must be used responsibly in a manner that is consistent with the individual's informed consent and reasonable expectations. To protect privacy, researchers should consistently implement systematic steps aligned with relevant laws and regulations and organizational policies and procedures to keep personal information about individual research participants confidential. Conversely, to improve the scientific utility of anonymized clinical data, it is important that researchers

[i] TransCelerate acknowledges that there are different definitions and understandings of pseudonymized and anonymized across the globe. In this paper anonymize is used to cover any measures taken to protect patient privacy, including pseudonymization allowing for a gradient approach as described in the International Standard ISO/IEC 27559:2022(E) [2].

[ii] State-of-the-art measures are described in the TransCelerate Biopharma Inc. paper "A Privacy Framework for Clinical Data Reuse: Secondary Data Use in the Pharmaceutical Industry." https://www.transceleratebiopharmainc.com/wp-content/uploads/2020/10/TransCelerate_A-Privacy-Framework-for-Clinical-Data-Reuse_Interpretations-of-Guidances-and-Regulations-Clinical_October-2020.pdf. October 2020.

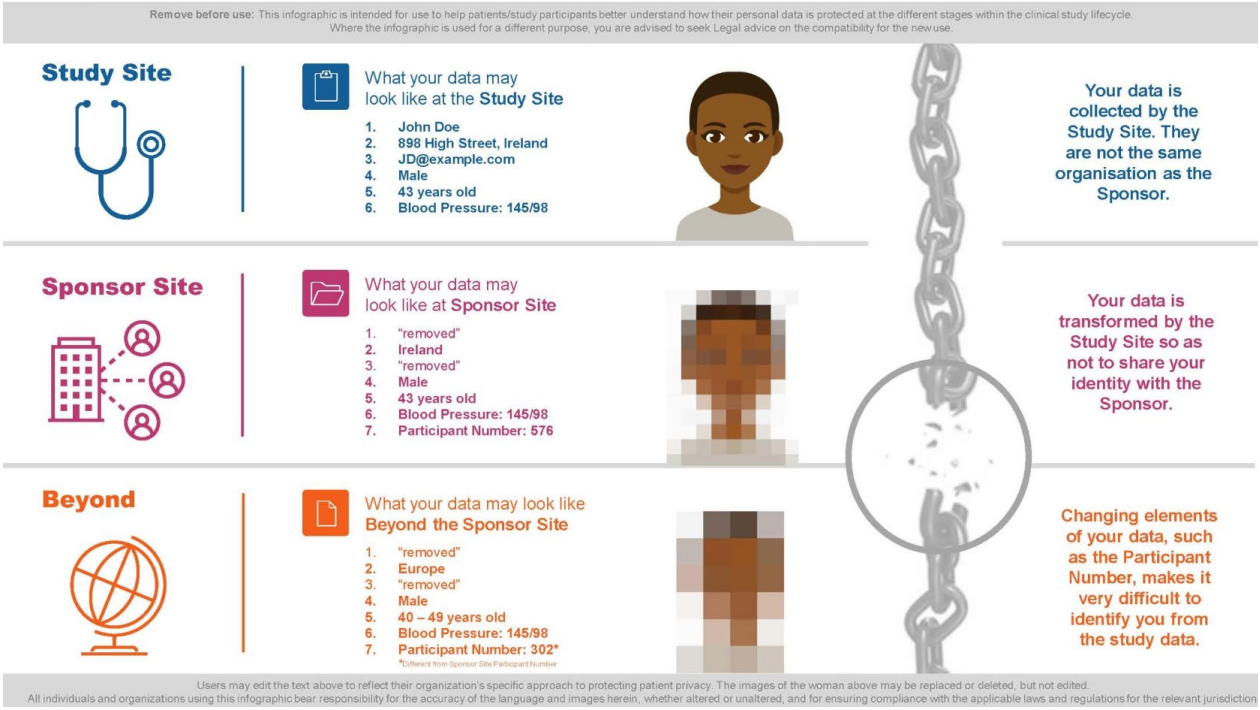
understand what steps have been applied to the data to ensure anonymization. These steps should be documented in sufficient depth and detail to provide contextual understanding of the original data to ensure that the anonymized data are appropriate for secondary research. At the same time, the level of documentation should not allow the researcher to re-identify the participants.

TransCelerate's 2016 publication "De-Identifying and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach" [3] included several best practices for sponsors; however, the data privacy approach taken across the pharmaceutical industry remains varied based on internal company policies.

TransCelerate is passionate about dynamic data and enabling the reuse of participant-level clinical trial data to improve the efficiency of clinical development and improve outcomes for patients. To that end, TransCelerate has established multiple data sharing initiatives (including exchange of clinical control arm data, COVID-19 data, and preclinical background control and toxicology data) and has developed the DataCelerate® platform to enable the secure and voluntary sharing of preclinical and clinical data. The proposed data privacy methodology outlined here aims to improve transparency and reduce the variability in the data privacy measures applied by companies in a way that enables improved secondary reuse of participant-level clinical trial data from DataCelerate®. Given that the same concerns faced by DataCelerate®, namely, the need to protect research participant privacy while facilitating data reuse, are also faced by other industry data sharing projects, elements of this proposed Data Privacy Methodology may be applicable to other clinical data sharing initiatives. The hope is that the methodology presented in this paper will also be useful for data sharing outside the scope of DataCelerate®.

This proposed methodology is the third deliverable created by TransCelerate to address issues related to privacy regulations and compliance in connection with secondary reuse of clinical trial data. The first deliverable was the 2020 framework "A Privacy Framework for Clinical Data Reuse: Secondary Data Use in the Pharmaceutical Industry" [4] to help companies address privacy questions related to clinical trial participant data and increase the potential reuse of clinical data. The second deliverable, released in 2022, was the "Educational Toolkit for Consent Specific to Data Reuse", [5] which provides Institutional Review Boards/International Ethics Committees, Health Authorities, and clinical trial participants with a customizable explanation of how anonymization works at a participant-friendly level. The infographic from the toolkit is presented below in **Figure 1**.

Figure 1. Infographic from TransCelerate's Education Toolkit for Consent



1.1 Overview of TransCelerate

TransCelerate BioPharma Inc. is a non-profit organization dedicated to improving the health of people around the world by accelerating and simplifying the research and development (R&D) of innovative new therapies. The organization's mission is to collaborate across the global biopharmaceutical R&D community to identify, prioritize, design, and facilitate implementation of solutions intended to drive the efficient, effective, and high-quality delivery of new medicines. The vast majority of TransCelerate solutions are publicly available. Headquartered in Philadelphia, TransCelerate has 22 member companies and 30+ initiatives focused on improving the patient and site experience, enhancing sponsor efficiencies and drug safety and, as appropriate, harmonizing process and sharing information.

Membership in TransCelerate is open to pharmaceutical and biotechnology companies with R&D operations. For more information, please visit http://www.transceleratebiopharmainc.com

1.2 Overview of DataCelerate®

DataCelerate® is a secure, United States Food and Drug Administration (US FDA) 21 CFR Part 11 compliant, cloud-based technology built on Accenture's Insights Platform hosted

by Amazon Web Services (AWS). The Placebo/Standard of Care (SoC) Module enables the sharing of participant-level control arm data from clinical trials. DataCelerate® supports several data sharing initiatives in the preclinical and clinical space via independent data repositories (i.e., modules). The clinical modules enable the sharing of participant-level control arm data from historical trials and participant-level COVID-19 data, respectively. To ensure the protection of personal data in DataCelerate®, physical and logical controls are built into the platform. Specifically, alongside established business processes, the member companies have data sharing and data processing agreements in place and all contributed data have been anonymized or pseudonymized[iii] to protect participants' privacy, in accordance with the rigor and quality employed by the participating pharmaceutical companies.

1.3 Target Audience

The target audience of this paper is pharmaceutical industry professionals who prepare clinical data for sharing or who utilize clinical data from external sources for research purposes. This may include, but is not limited to, data managers, transparency and disclosure experts, data scientists, biostatisticians, translational researchers, statistical programmers, and legal and compliance professionals.

1.4 Use of Methodology

TransCelerate did not develop this methodology as a replacement for current anonymization approaches used by sponsors, vendors, and other relevant stakeholders. In fact, application of the methodology, without more, may not be sufficient to anonymize clinical trial data. Users should consider how the methodology works in connection with their current approaches to increase data utility. Users of the methodology remain responsible for their own compliance with all applicable laws and regulations.

2.0 BENEFITS FOR DATA PROVIDERS AND DATA USERS

The proposed data privacy methodology presented here focuses on clinical trial datasets shared among the TransCelerate member companies in the DataCelerate® database. As such, the recommended approach is based on specific conditions of that project. However, the hope is that the methodology may also be useful for data sharing outside the scope of DataCelerate®.

[iii] Which may include HIPAA 18-factor deidentification, pseudonymization, or anonymization as defined by the European Union (EU), as appropriate.

Certain variations in the data privacy approach used across companies or studies can make the resulting datasets time-consuming and difficult to understand and use, and in some cases, can greatly reduce the potential for cross-study analyses. The challenge is further complicated when there is a lack of transparency in describing the specific data privacy safeguards that have been applied, making it difficult for a Data User to determine whether the data are fit for an intended secondary purpose.

Building on existing solutions in the industry, such as those developed by PHUSE—The Global Healthcare Data Science Community [6], TransCelerate [7], and the European Union Agency for Cybersecurity (ENISA) [8], the proposed data privacy methodology outlined in this paper recommends data modification approaches for relevant data variables typically found in clinical datasets.

Ultimately, the proposed data privacy methodology is intended to:

- Preserve privacy and confidentiality of research participants.
- Reduce research participant burden through easier reuse of existing study data.
- Deliver faster scientific insights.
- Provide greater transparency in the privacy safeguards applied to study data.
- Increase overall data utility.

Data Providers benefit from this methodology by having systematic guidance for identifying critical privacy variables to protect and for how to appropriately apply these safeguards. If uniformly applied, this approach is one step towards improving data consistency that better enables cross-study analyses.

Data Users benefit from this methodology by having greater transparency and confidence in the data they are receiving from Data Providers who follow this data privacy approach.

3.0 ASSESSMENT OF RE-IDENTIFICATION RISK: QUANTITATIVE OR QUALITATIVE APPROACHES

To adequately protect the individual research participant's privacy and preserve the integrity of their data, it is critical to consider the effectiveness of the anonymization process. Even when removing direct identifiers (such as name or unique subject ID), there is a risk that the participant's identity can still be revealed through a single or combined piece(s) of personal information left in the dataset (e.g., rare genetic mutation), single study site in an isolated geographic location) or when data in the dataset is compared to or combined with other data sources. Assessment of this re-identification risk can be

performed using many different methods but, in general, the methods can be categorized as a quantitative or qualitative approach.

Approaches for de-identifying/anonymizing clinical data have evolved over time. In TransCelerate's model approach from 2016 [3], there is a description of the "HIPAA: Safe Harbor Method" (today generally considered an insufficient approach across the pharmaceutical industry) and the "Expert Determination Method" (similar to the qualitative approach). Qualitative re-identification risk assessment relies on subject matter experts determining the probability and impact of re-identification based on experience and knowledge. After the implementation of the European Medicines Agency (EMA) Policy 0070 [9] and other similar requirements, many companies now also rely on a quantitative approach, relying on objective, measurable data to provide insights into the re-identification risk against a pre-determined "risk threshold".

Both quantitative and qualitative approaches or a combination of both, if applied systematically, will help determine re-identification risks and guide organizations on what information needs to be removed, changed, or redacted to protect the identity of an individual before sharing study data. This document is not intended to provide guidance on how to measure the risk of re-identification nor set risk thresholds for specific types of clinical data. TransCelerate acknowledges that the risk threshold may differ depending on other factors not linked to the individual dataset, for example the environment in which data is shared. For this reason, organizations should, when deciding on a risk threshold, also consider which legal, organizational, and technical safeguards are in place. Sharing data in a closed environment with a limited number of researchers for a specific purpose will involve a different risk than sharing data via publicly available Health Authority portals. Organizations may need to consider how many different places or purposes a dataset is shared since it can increase the likelihood of re-identification by comparing different versions of the dataset, such as:

- Applying different risk thresholds across multiple data sharing environments.
- Targeting anonymization to individual research proposals.

Following the EMA Policy 0070 [9], it is our recommendation that a quantitative, re-identification risk assessment approach is followed and that risk thresholds are set consistently across studies where the re-identification risk is considered similar (e.g., same risk threshold could be defined for studies in the same rare disease).

By aligning how variables are transformed during data protection steps and documenting method differences in supporting documentation, it is possible to achieve transparency, repeatability, consistency, and provenance[iv] when using clinical data for secondary research.

[iv] *Provenance is described as an audit trail that explains the origin of data, and a justification of how it arrived at its present place, which provides confidence in the integrity of the source data.*

4.0 PROPOSED PRIVACY METHODOLOGY

Data Providers ultimately must decide what they consider adequate approaches to data protection to comply with relevant laws, regulations, and commitments made to patients. Moreover, some companies may already have a preferred internal approach to data protection or are using third-party products or suppliers to ensure adequate privacy that are not easily changed.

The aim of this proposed methodology is to highlight variables where creating consistency should improve data utility and facilitate cross-study analysis. The following sections describe primary data variable types where TransCelerate member companies use privacy preserving techniques and approaches to reduce the risk of re-identification of research participants from clinical data. For each element/variable, we propose a **recommended approach** and a **compatible approach**. By providing two approaches, we strive to increase the usability of this methodology while maintaining a level of uniformity for data privacy and improving data utility.

In general, the recommended approach should provide increased scientific utility of the data, whereas the compatible approach provides an alternative, a more conservative method that may not lead to the highest data utility but may have a preferable risk-benefit aligned with the Data Provider's policies. For increased transparency, it is good practice to provide as much information as possible on the data protection approaches taken in the supporting documentation (**Section 6.0 Supporting Documentation**).

TransCelerate is aware that laws and regulations impact the approach taken to anonymize the datasets. **Please note, each company remains responsible for ensuring the privacy and security of the contributed datasets is in accordance with their internal procedures and defined risk thresholds as well as relevant laws and regulations.** Our recommendations are not intended to supersede an organization's compliance obligations.

4.1 Study Unique Identifiers

During a clinical trial there may be a significant amount of collected data or events that potentially could uniquely identify a study participant, either on their own or in combination with other data values, and thus present a re-identification risk. Hence, the identification potential of all variables within a structured clinical dataset (i.e., tabular format) should be assessed on a study-by-study basis to determine if they are likely to provide values that uniquely link to a study participant.

These study unique identifiers can be split into two categories:

1. **Direct Identifiers**, i.e., data values that, on their own make it possible to uniquely identify a study participant. Direct identifiers could include Unique Subject ID (USUBJID), serial number or outliers.
2. **Indirect Identifiers** (also known as quasi-identifiers), i.e., data values that, in combination with others in the dataset make it possible to uniquely identify a study participant. Indirect identifiers could include demographic information, site identifier, vendor identifier, batch/lot numbers or study administrative data variables.

Non-identifiers are another category which may exist within a clinical dataset. Non-identifiers are data variables or values that are either not linked to an individual trial participant or do not contribute to participant re-identification. Examples include trial code/study ID, EUDRA or IND number, NCT number.

4.1.1 Recommended Approach

For direct identifiers, the recommended approach is to replace the original values with scrambled values of the same length, type, and format to enable consistency between datasets and document de-identification/anonymization. Please see Table 1 for an example.

For indirect identifiers, the recommended approach is to assess if anonymization is needed on a study-by-study basis. If it is not needed, the indirect identifier should be retained (as exemplified in Table 1 for xxSEQ). If it is not possible to retain the indirect identifiers, they should be scrambled rather than redacted/suppressed, if possible.

For example, derived identifiers can be retained if they are only for internal study administrative purposes and assessed to present no re-identification risk either on their own or combined with other data.

Variables or unique values classified as non-identifiers should be retained.

Table 1. Example of scrambling study unique identifiers

	BEFORE	AFTER	BEFORE	AFTER	NO CHANGE[v]
STUDYID	**USUBJID**	**USUBJID**	**SPDEVID**	**SPDEVID**	**xxSEQ**
1234-5678	1234-5678-10001	1000-0010-00056	JBDDCG0-1011	AAAAAAA-0110	1
1234-5678	1234-5678-10001	1000-0010-00056	JBDDCG0-1011	AAAAAAA-0110	2

[v] xxSEQ is considered a non-identifier and as such left as-is.

1234-5678	1234-5678-10002	1000-0010-00301	JBDDCG0-1013	AAAAAAA-0074	4

4.1.2 Considerations

Studying unique identifiers will increase the likelihood of re-identification, either by being unique to an individual participant or by potentially providing a link to additional information about the trial via other sources (e.g., trial registries). Across the industry there is a common understanding that direct identifiers should not be left in the dataset. PHUSE [10] guidance points to the removal of direct identifiers, except for the Investigator Identifier, if this is needed for investigator analysis. However, pseudonymized direct identifiers are often needed as a key to link data values to a specific subject or when comparing different values, they can also be relevant for further research. Therefore, it is recommended to scramble rather than remove the direct identifiers to enable this type of use. This may not be possible for all data variables (e.g., outliers, which will likely need to be removed from the dataset).

As with other identifiers, an assessment of re-identification risk should be performed for study unique indirect identifiers. The risk associated with retaining the study unique indirect identifier versus the relative utility of the indirect identifier should be considered after other indirect identifiers have been transformed. It is important to also consider if from a scientific or medical perspective there could be any specific requirements that necessitate the retention of indirect identifiers. To the extent possible indirect study unique identifiers should be retained. If the risk or re-identification is too high, then the indirect identifier should be scrambled rather than removed.

4.1.3 Compatible Approach

The compatible approach is to redact, remove or overwrite all unique identifiers except the Participant ID (e.g., USUBJID in **Table 1**), which is considered a primary key variable and must be maintained in scrambled form. The rationale for redaction should be described in the supporting documentation.

4.2 Dates

Dates within clinical data can provide scientific utility but could also be used to re-identify the individual participants (e.g., looking at the pattern and cluster of information that might be retrieved by a set of dates in conjunction with separate data sources). To protect the research participant's privacy, companies tend to transform dates related to individuals.

You may already have a mix of relative days and dates within your study. Relative days are generally not identifying so they may be left as is. If actual dates exist in your dataset,

then they should be transformed. The two most common methods of transforming dates are Relative Day and Date Offset.

Relative Day is a method that uses a specific date (e.g., each participant's randomization date) as a reference day (e.g., day 0, day 1) for a participant, and transforms all other days for that specific participant into the number of days relative to the reference date. The original date variables are to be redacted afterward.

Date Offset is a method where the original date is transformed by adding or subtracting a defined number of days from the original date. This method can be used with different date ranges and includes an option to use a different randomized offset on an individual basis.

4.2.1 Recommended Approach

It is recommended that if dates are not redacted, they should be transformed using the Date Offset method with a randomly generated number of days (within an appropriate range for the study, Table 2). This number of days is added to or subtracted from each research participant's dates (i.e., the offset is applied on a participant level, not one offset applied across the whole study). For legacy datasets where different date formats may exist, the recommendation is that date formats should, where possible, follow CDISC standards (e.g., YYYY-MM-DD) after transformation to improve cross-study analysis when multiple studies are analyzed.

Table 2. Example of Date Offset

		BEFORE	AFTER	BEFORE	AFTER	DATE OFFSET[vi]
STUDYID	**SUBJID**	**DATE1**	**DATE1**	**DATE2**	**DATE2**	
1234-5678	1234-5678-10001	2021-03-02	2021-03-24	2022-07-08	2022-07-30	22
1234-5678	1234-5678-10002	2021-08-29	2021-09-24	2022-07-31	2022-08-26	26
1234-5678	1234-5678-10003	2021-11-06	2021-11-01	2022-07-29	2022-07-24	-5

[vi] The offset number per participant should not be shared.

4.2.2 Considerations

Dates are collected in clinical data for many different purposes. However, not all dates are considered sensitive information or pose a risk to re-identification (e.g., study-related dates). Dates that are related to individual participants (e.g., event/visit dates) should be considered indirect identifiers because they could be used alongside other information to re-identify individuals. Dates that are not related to an individual participant (e.g., medication batch dates or study start and end date) typically do not require transformation since they are unlikely to be associated with any re-identification risks.

When using the Date Offset method, the selection of most appropriate offset range to apply within a study (e.g., ±30 days, ±90 days) can be study dependent. In cases with a high proportion of dates within a small range, a small offset range might be sufficient. Meanwhile, in studies with a prolonged enrollment period, broader offset ranges might be needed to protect the participant's privacy.

Offsetting partial or imputed dates requires special attention because this method may reveal the number of days used to create the offset dates, and thus expose the original dates. For example, imputation in clinical trials often follows standard rules (i.e., missing day is imputed by 01, missing month by 01 or 06), and these rules are often explained in Statistical Analysis Plans or other supporting documentation, which could be made public. Therefore, there is a risk of knowing that a value has been imputed and, by comparing the imputed value to the de-identified/anonymized value, the offset range could be revealed.

To minimize the risk and keep high data utility, it is recommended to remove all information associated with the imputed date values, i.e., imputation flags (CDISC ADaM *DTF variables) and incomplete source variables.

For studies where seasonality is important, the company should judge if a narrow band might be applied to keep the seasonal effect. Minimum and maximum offset values can be important to include in the supporting documentation (**Section 5.2 Seasonality**).

4.2.3 Compatible Approach

An alternative compatible approach is to have Relative Day calculation together with date redaction, where all dates related to the research participants have been redacted.

4.3 Verbatim/Free Text

Verbatim text (otherwise known as "free text") may be collected across a wide range of case report form (CRF) pages and thus may be present in many datasets. Since any text strings could be captured within these fields, it is likely that personal data are collected.

In some instances, these fields are used for coding purposes (e.g., adverse events are originally captured as verbatim text that may be subsequently coded using the Medical Dictionary for Regulatory Activities [MedDRA]). For analysis purposes, the coded variables are then utilized to produce analyses such as frequency tables.

Other verbatim variables may provide valuable insight or information to other data fields collected, though their lack of structure can be difficult to analyze, and a value-by-value examination may be needed for partial redaction.

4.3.1 Recommended Approach

In general, free text fields should be redacted with some non-blank value(s) to avoid any ambiguity whether the variable or column was redacted or originally blank.

For any verbatim/free text fields in the study datasets that require modification, the field or specific wording should be replaced with the string "—redacted—" (Table 3).

Table 3. Example of Redaction of Verbatim/Free Text

		BEFORE	AFTER
STUDYID	**SUBJID**	**COVAL**	**COVAL**
1234-5678	1234-5678-10001	Comment 1	—redacted—
1234-5678	1234-5678-10002	Comment 2	—redacted—
1234-5678	1234-5678-10003	Comment 3	—redacted—

4.3.2 Considerations

There is a variety of information collected within a study that could influence the degree of privacy risk associated with free text, such as use of unusual names, diseases, medication (Section 4.9 Concomitant Medications), or rare adverse events (Section 4.8 Adverse Events and Medical History). Information on these elements may present a re-identification risk. However, the use of free text can vary greatly between studies and organizations. Establishing an effective balance between preparation effort, data utility, and privacy risk needs to be considered within the data transformation process.

Some organizations use tools to electronically scan free text for sensitive terms. Depending on the number of verbatim/free text fields within in a study database, a manual review of all free text fields for personal information may be resource intensive.

If a sensitive term is identified, it should be removed based on the Data Provider's policies and practices and replaced with a term such as "—redacted—" (as recommended by PHUSE [10]). If information cannot be retained in a dataset, it is preferable for a Data User

to know that information has been removed. The replacement of free text with a term such as "—redacted—" is therefore preferable to setting the field to blank.

4.3.3 Compatible Approach

If a Data Provider protects personal data by setting fields to blank rather than replacing it with a term such as "—redacted—," the compatible approach requires the following additional information to ensure data utility:

- Comprehensive information in the supporting documents to provide secondary research teams with details as to which variable/columns were blanked out to avoid ambiguity, and/or
- Derived information/variables provided in substitute of the removed information.

4.4 Banding of Variables

Unlike direct identifiers, many indirect identifiers are critical for data utility. Indirect identifiers, such as age, height, weight, and body mass index (BMI), are considered to have high scientific utility but may also pose a risk for re-identification. The proposed approaches here transform such variables to reduce the risk of re-identification while maintaining a level of data utility.

Banding is a commonly used technique where a continuous variable such as age is aggregated into bands. During this process, a specific number is transformed into a range, thereby reducing the likelihood that a participant may be singled out in the data.

The appropriate band size (value range) to use will be one that strikes a balance between factors that allow a narrower band size to maximize data utility (e.g., purpose of the intended research) and factors that require a broader band size to lower the risk of re-identification (e.g., the distribution of the trial population on the variable to band, such as age; the presence of outliers, or unique values, which could also be treated as low frequency data (see **Section 4.6 Data with Low Frequencies**)).

Once the band size is determined, it can be applied using different banding techniques that exist:

- **Static bands:** Static bands have the same cut-off points for bands for each study (e.g., 20–29, 30–39)
- **Semi-fixed bands:** Semi-fixed bands support the combination of selected static bands to increase the number of participants within a band to reduce privacy risks (e.g., 20–39 when combining bands from static banding example above)
- **Flexible bands:** Flexible bands are individual bands created for each set of data to preserve scientific utility as much as possible. Flexible bands come in two varieties: single-dimensional banding, which generates independent bands on each relevant variable within the dataset, and multi-dimensional banding, which

consists of creating bands on two or more variables (e.g., based on age and some prior medical history diagnosis)

4.4.1 Recommended Approach

Single-dimensional flexible banding is recommended, with bands as small as possible. If band covers only 1 year (e.g., 50), then the recommendation is to specify this as "year-year" (e.g., 50–50) to demonstrate a conscious modification has been applied to the variable (Table 4).

Table 4. Examples of Single-Dimensional Flexible Banding of Age and Height

		BEFORE	AFTER	BEFORE	AFTER
STUDYID	**USUBJID**	**AGE**	**AGE**	**HEIGHT**	**HEIGHT**
1234-5678	1234-5678-10001	46	46-48	153	150-159
1234-5678	1234-5678-10002	47	46-48	161	160-169
1234-5678	1234-5678-10003	48	46-48	175	170-179
1234-5678	1234-5678-10004	50	50-50	168	160-169
1234-5678	1234-5678-10005	50	50-50	163	160-169
1234-5678	1234-5678-10006	50	50-50	180	170-179
1234-5678	1234-5678-10007	50	50-50	210	≥190

4.4.2 Considerations

In most cases, not applying a transformation such as banding to indirect numeric identifiers may represent an unacceptable re-identification risk. Banding is a method already widely adopted among companies to reduce this risk as part of their qualitative or quantitative risk approaches.

Static bands are not recommended as these would need to be sufficiently large to cover a wide range of studies, which would be detrimental to data utility.

Semi-fixed banding and flexible banding would require application of the largest bands when combining data. Semi-fixed banding, however, may be preferable, as the number of banding edges are fewer and therefore may reduce the variability of the number of bandings, which may lead to improved consistency across variable types.

Multi-dimensional banding techniques are not yet widely adopted, and while promising a significant increase in data utility, might not be able to be adopted quickly. Multi-dimensional banding also might impact the work of statisticians who are used to working

with single-dimensional banding. For this reason, we do not recommend multi-dimensional banding.

In case a single number is desired instead of a range, specific numbers can be replaced with terms such as "random number in band" or "mean within band." If this approach is used, care should be taken to make sure this approach is appropriately documented for transparency.

4.4.3 Compatible Approach

Semi-fixed banding is an alternative approach to flexible banding but may result in lower overall data utility.

4.5 Patient Demographics (Sex, Race, Ethnicity)

Demographic data collected as part of the clinical study provide an essential element to describe the study population as part of the planned analyses. Information such as sex, race, ethnicity, and age are valuable to retain for data sharing and further data reuse. However, combined and correlated together with other quasi-identifying information, the overall risk of re-identification might increase if no further measures are applied.

Sharing the demographic data of trial participants improves the quality of analysis in several respects. Demographic data can be aggregated and analyzed to better understand disease states and treatment effectiveness among diverse populations, and the insights generated from such analyses can be prospectively incorporated into future clinical studies to advance inclusive research. The desire to share as much demographic data as possible must be balanced with the need to protect participant privacy by redacting and/or grouping some of the demographic data.

4.5.1 Recommended Approach

The recommendation is to retain as much information as possible about sex, race, and ethnicity.

For low frequency events, the recommendations in this section should be considered with **Section 4.6 Data with Low Frequencies** and **Section 4.7 Sensitive Information**. The privacy risk must be assessed to determine the re-identification risk in conjunction with other indirect identifiers (e.g., age, height, weight). Where the risk proves to be significant (according to the risk threshold), information about the low frequency event(s) should be removed.

In cases where the demographic attributes cannot be retained, additional information revealing the research participants' sex (e.g., info about pregnancy tests, erectile dysfunction) or race/ethnicity (e.g., regional relationship) should be removed as well.

When information related to a participant's sex, race, and ethnicity and other indicative attributes have been removed, the rationale should be included in the documentation supporting any data sharing package.

4.5.2 Considerations

Assuming a well-balanced study population and following re-identification risk assessments, most Data Providers retain information about sex, race, and ethnicity in the dataset, to the extent that no low frequency groups are present in the data.

Outliers are interesting and important factors in the analyses but must be given special consideration because there may be an increased risk of re-identification for the specific participant.

Considerations specific to re-identification of non-binary research participants' sex (CDISC (Clinical Data Interchange Standards Consortium) controlled terminology: "BOTH"/ "Unknown"/ "Undifferentiated") are required due to the relative rarity.

The De-identification Standard for Study Data Tabulation Model (SDTM) 3.2 [10] developed by the PHUSE Data Transparency Working Group, references the FDA (Food and Drug Administration) guidance document "Collection of Race and Ethnicity Data in Clinical Trials" (2016) [11] and further makes recommendations to consider "race" in relation to geographic region in case no race/ethnicity information is collected for a certain country.

4.5.3 Compatible Approach

No alternative approach is proposed.

4.6 Data with Low Frequencies

Variables that contain some low frequency values may lead to an increased risk of re-identification (e.g., data that, after grouping of several variables and their expression levels, applies to a very small cell/group size). This may be further compounded by data with multiple variables of low frequencies.

4.6.1 Recommended Approach

It is recommended that variables with low frequencies should be assessed, and if there is a significant risk of re-identification, those values with low frequencies should be redacted (**Table 5** and **Table 6**).

The De-identification Standard for SDTM 3.2 document from PHUSE [10] serves as a guide to make the appropriate adjustments to the data. According to PHUSE, where data with low frequencies is shared to a controlled environment, such as DataCelerate®, an average risk might be assumed and "(…) a minimal frequency of 5 (or probability of

0.2) may be acceptable for a secure portal data release under certain conditions (...)" [10]. Therefore, the specific recommendation is to apply a company-determined minimum frequency value (e.g., 5) for a trusted data disclosure.

Removal or redaction of the low frequency event(s) should be explained in supporting documents.

Table 5. Example of Redacting Low Frequency Sex and Race

		BEFORE	AFTER	BEFORE	AFTER
USUBJID	**DOMAIN**	**SEX**	**SEX**	**RACE**	**RACE**
1234-5678-USA003-10001	DM	F	-- Redacted --	WHITE	WHITE
1234-5678-DNK001-10002	DM	F	-- Redacted --	WHITE	WHITE
1234-5678-POL002-10003	DM	M	-- Redacted --	WHITE	WHITE
1234-5678-GER002-10004	DM	F	-- Redacted --	UNKNOWN	-- Redacted --

Table 6. Example of Redacting Additional Information Revealing Trial Participants' Sex Through a Rare Event (e.g., Event of Oligospermia)

		BEFORE	AFTER
USUBJID	**DOMAIN**	**AEDECOD**	**AEDECOD**
1234-5678-POL002-10003	AE	Oligospermia	-- Redacted --
1234-5678-POL002-10003	AE	Headache	Headache
1234-5678-POL002-10003	AE	Diarrhoea	Diarrhoea
1234-5678-POL002-10003	AE	Rhinitis	Rhinitis

4.6.2 Considerations

A great deal can be learned from certain low-frequency study data so there is usually a special interest to analyze rare events and outliers in a clinical study. Careful consideration should be given to whether these study data should be retained as the risk of re-identification significantly increases when participant data falls into a very low frequency group.

For a deeper dive into the topic and for additional guidance, Appendix 2 "Low Frequencies" of the PHUSE Data De-Identification Standard for SDTM 3 [10] is recommended.

4.6.3 Compatible Approach

No alternative approach is proposed.

4.7 Sensitive Information

Sensitive information[vii] is highly personal in nature and disclosure may cause harm to the individual participant (e.g., data related to alcohol abuse, drug use, conditions such as HIV/AIDS, mental health information).

4.7.1 Recommended Approach

It is recommended to redact sensitive information across all variables (Table 7). The redaction of sensitive information should be explained in supporting documents.

It is also recommended that companies have a policy to determine what variables should be classified as sensitive information, as that can vary across therapeutic areas.

Table 7. Example of Removing Sensitive Substance Usage Records

			BEFORE	AFTER
USUBJID	**DOMAIN**	**SUCAT**	**SUOCCUR**	**SUOCCUR**
1234-5678-USA003-10001	SU (Substance Usage)	ALCOHOL HISTORY	N	-- Redacted --
1234-5678-USA003-10001	SU	TOBACCO HISTORY	Y	-- Redacted --
1234-5678-GER002-10004	SU	ALCOHOL HISTORY	Y	-- Redacted --
1234-5678-GER002-10004	SU	TOBACCO HISTORY	Y	-- Redacted --

[vii] *By sensitive information this document leans on the definition of sensitive data provided in the EU GDPR. However, as clinical data by definition refers to health data, what is considered sensitive information in a clinical dataset will depend on the context and the therapeutic area which the dataset relates to. As an example, data referring to HIV would not be considered sensitive information (as described in this document) in a dataset related to the study of HIV/AIDS but may be considered sensitive information in a dataset related to the study of Obesity.*

4.7.2 Considerations

Depending on the clinical research, some studies will inherently involve the capture and analysis of sensitive information to determine the primary or secondary outcome measures of the study. Sharing this sensitive data has a much greater risk of causing harm to the affected research participants in the event of re-identification. Given the severity of the potential harm, additional measures (e.g., safeguards in information security, user access management, and specific use case limitation) must be factored in the re-identification risk assessment when sharing sensitive information.

Sensitive information could also relate to some variables already mentioned in this proposed methodology, such as patient demographics. Depending on the trial's areas of study, those variables should be assessed following the recommendations in this section.

Redaction is preferrable to data removal (e.g., blank fields) to ensure that Data Users can distinguish when a data field has been transformed/modified versus when the field was missing in the original dataset.

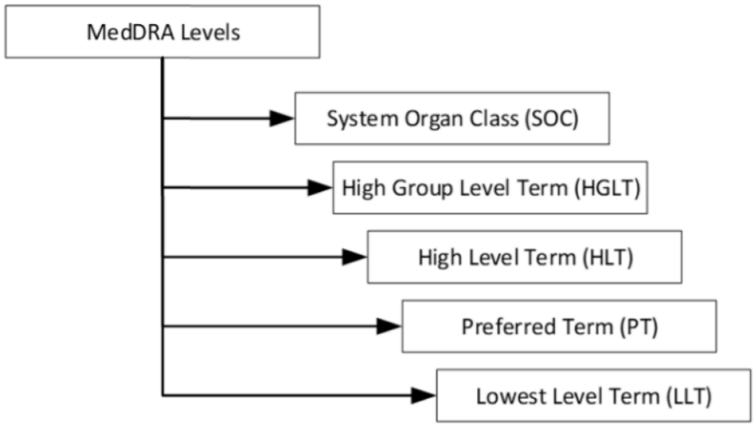
4.7.3 Compatible Approach

No alternative approach is proposed.

4.8 Adverse Events & Medical History

Collection of adverse events during the clinical study plays a pivotal role in assessing the safety of the investigational product. Similarly, understanding the medical history and comorbidities of a research participant can help assess safety and efficacy. The MedDRA dictionary provides codes to describe the adverse events terms (AE) at 5 levels: System Organ Class (SOC), High-Level Group Term (HLGT), High Level Term (HLT), Preferred Term (PT), and Lowest Level Term (LLT) (**Figure 2**). The same dictionary level is typically used for documenting medical history. Values related to the medical history date or concomitant medication date, should follow the recommendation on dates, see **Section 4.2 Dates**.

Figure 2. MedDRA Dictionary Levels



4.8.1 Recommended Approach

It is recommended that all 5 levels of the MedDRA dictionary are shared along with the version of the MedDRA dictionary being used. This information should be included in the supporting document (see **Section 6.0 Supporting Documents**).

Further, it is recommended that the verbatim adverse event terms (AETERM) are redacted and replaced with the corresponding MedDRA coded term (**Section 4.3 Verbatim/Free Text**).

Following appropriate processes, if a Data Provider identifies rare medical history or adverse events that pose re-identification risk, then the information should be redacted to protect participant privacy. The applied redaction steps should be described in the accompanying supporting documentation so that potential Data Users could assess whether the study is suitable for their secondary research purpose.

4.8.2 Considerations

The more levels of MedDRA that are shared with respect to the adverse event or medical history, the greater the level of utility that can be ascertained from the information. However, the more granular the description of the adverse event, the greater the possibility that a participant could be re-identified using these data or in combination with other data pertaining to them.

Consideration needs to be made regarding the nature of the adverse event and whether other information, when combined, increases the identifiability of a specific participant.

As Data Providers are preparing adverse events for data sharing, they should consider how common an adverse event or medical history is as this can impact the risk of re-identification. Very common adverse events would be considered lower risk and more

rare AEs would be considered higher risk for re-identification. This should be accounted for when considering if all 5 MedDRA code levels can be shared.

4.8.3 Compatible Approach

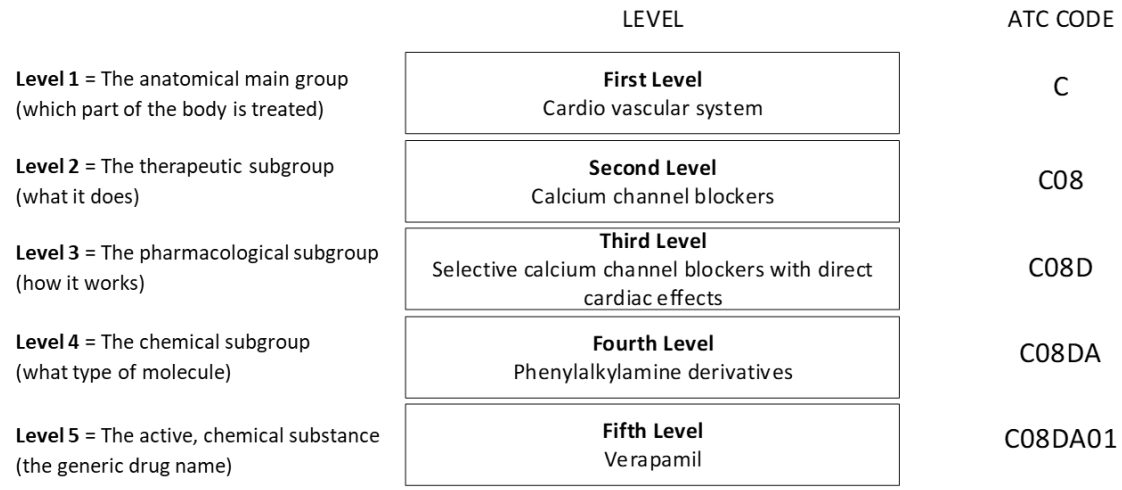
Where the risk of re-identification within the study has been assessed as too high to share all 5 MedDRA levels, it is acceptable to share fewer levels (e.g., not share LLT). This lowers the privacy risk to the individual but may also reduce the study data utility. However, as standard adverse event or medical history analyses are typically performed at the SOC/PT level, sharing of LLT may lead to minimal additional utility.

4.9 Concomitant Medications

Information collected during clinical studies includes the medication history of research participants. This consists of the current and concurrent medications that the participant is taking at the time of the study as well as medications taken in the past. Accurate documentation of this information is critical for researchers to understand whether a participant's medication history should be treated as a confounding factor in the original clinical study or in subsequent analyses.

Clear and consistent identification of the medications taken by study participants is essential to maintaining data utility so that Data Users can correctly assess what impact, if any, the participants' medication use may have on the resulting study data. To help achieve consistent communication of medications, the World Health Organization (WHO) established a structure for medication classification, known as Anatomical Therapeutic Chemical (ATC) Classification (**Figure 3**) [12].

Figure 3. ATC Classifications



Source: Verapamil example from WHO website

4.9.1 Recommended Approach

Provided the Data Provider has the appropriate WHO ATC code license, it is recommended that the complete WHO ATC code (all 5 levels of detail) is shared for the dataset along with the version of the WHO ATC classification being used. **The version information should be included in the supporting documents unless it is provided in a variable in the dataset.**

Where knowing the medication used by a participant is considered to increase the risk of re-identification, it is acceptable to share a partial code containing fewer levels (e.g., only first to fourth level or first to third level). This may lower the privacy risk to the individual but may also reduce the data utility. To help Data Users, the Data Provider should describe the reasons behind the reduced code in the supporting documents and the WHO ATC code version used.

It is also recommended that the verbatim medication term (CMTRT/CMMODIFY) is removed and replaced with the corresponding WHO ATC drug code (**Section 4.3 Verbatim/Free Text**).

4.9.2 Considerations

Licenses – According to WHO Uppsala Monitoring Centre (Uppsala Monitoring Centre | UMC (who-umc.org)), the WHO drug code used by the contributing organization(s) works with a subscription-based license. The license requires all companies intending on using the WHO drug code to have an in-date license. Therefore, to share the ATC code, the Data Provider and Data User would need to have a valid license. For more

information about license (see **Section 4.12 Information Collected Under Copyright Licenses**).

Version – the WHO ATC drug coding is periodically updated. Thus, it is important to document and share which WHO classification version number is used for the dataset (**Section 6.0 Supporting Documents**).

For instances where a particular medication is not frequently used, or when the use of a particular medication allows the inference of other information that has been suppressed, sharing the entire ATC code (all levels) may increase the possibility of re-identifying a participant. Some examples include when the brand name of medications can indicate geographic location, when medication with controlled use may be sensitive, and when medications may indirectly reveal sensitive conditions (e.g., HIV, mental illness, rare diseases).

4.9.3 Compatible Approach

No alternative approach is proposed.

4.10 Geographic Location

Information related to geographic location can serve as indirect identifiers that may increase the risk of re-identification when combined with other available information in participant-level data. When associated with an individual participant's data, location information that is specific to the study could allow for links between data and facilitate the recreation of an individual participant's profile.

4.10.1 Recommended Approach

The recommendation is to retain the original country information, when possible, or to keep it as close as possible to the data collected. Generalizing geographic information to the continent level by aggregating with other countries on the same continent is only recommended in cases of single-site countries and/or countries with very few participants. In these cases, generalizing the country information to the region or continent will maintain the protections to participant while preserving sufficient data utility to support analyses by region or by country (**Table 8**).

It is recommended to apply the methodology proposed by the United Nations Statistics Division (UNSD) by using the ISO-alpha3 and M49 (UNSD — Methodology) [13] standards for the generalization of geographic information. Avoid using "other" as a geographic location category.

Please note, other identifiers could indirectly point to geographic location, such as any study site identifiers or related information (e.g., address, phone number). These identifiers would also need to be modified.

Table 8. Example of Retaining Country Information for USA and DNK While Generalizing Country Information to the Continent Level for Poland and Germany (Only One Site per Country)

	BEFORE	AFTER	BEFORE	AFTER	AFTER
SUBJID	**SITEID**	**SITEID**	**COUNTRY**	**COUNTRY**	**REGION**
10001	USA-0001	011-0110	USA	USA	NORTHERN AMERICA
10002	USA-0001	011-0110	USA	USA	NORTHERN AMERICA
10015	USA-0002	011-0221	USA	USA	NORTHERN AMERICA
10037	POL-0001	054-0080	POL	-- Redacted --	EUROPE
10042	GER-0001	036-0073	GER	-- Redacted --	EUROPE
10068	DNK-0001	088-0004	DNK	DNK	EUROPE
10070	DNK-0002	088-0013	DNK	DNK	EUROPE

4.10.2 Considerations

In general, retaining geographic location helps maximize data utility, particularly when pooling multiple studies. However, in the case where there are very few study sites and/or participants per country, geographic information could be used to re-identify individual participants. To reduce this risk, generalizing location to the continent level provides a variable with less geographic specificity and therefore would not be as useful for re-identifying research participants. Nonetheless, this approach would still allow some geographic information to be retained.

The use of "other" or similar nondescriptive wording to replace geographic location introduces unwanted ambiguity to study data. As the composition of "other" cannot be discerned, this ambiguity becomes an issue, particularly when countries with very different healthcare systems/populations are combined.

In a therapeutic area where seasonality is relevant and was a component of primary analyses, include an additional variable indicating the hemisphere (northern, southern, equator). See **Section 5.2 Seasonality** for considerations. The Data Provider will need to assess and adjust the anonymization approach accordingly to ensure this approach does not increase the risk of re-identification.

4.10.3 Compatible Approach

No alternative approach is proposed.

4.11 Records of Participants Who Have Died

Some research participants may die during the conduct or post-treatment phases of the study. In some countries, privacy regulations may apply post-mortem (e.g., for 10 years after death). While records of participants who have died is of analytical value, care must be taken to recognize the personal nature of this information and the need to show courtesy to the next of kin.

The Data Provider is responsible to ensure that use of data from participants who have died is compliant with any applicable regulations.

4.11.1 Recommended Approach

The recommendation is to treat records of research participants who have died in the same manner as records of a living person, affording the participant all appropriate privacy considerations.

A participant's death should have no special impact on the level of anonymization needed for that individual and there are no specific additional steps needed for participants who have died when applying this methodology. Therefore, the same data privacy measures applied to living research participants' data should be applied to data from research participants who have died.

4.11.2 Considerations

The loss of children or young adults may be particularly distressing for the next of kin; safeguarding their privacy is therefore paramount.

Some privacy regulations, such as the General Data Protection Regulation (GDPR) [14] and the UK Data Protection Act [15], no longer apply to identifiable data that relate to a person once deceased. However, any duty of confidence established prior to death does extend beyond death so it is important to maintain privacy and confidentiality to ensure that trust in services and institutions are not undermined.

4.11.3 Compatible Approach

No alternative approach is proposed.

4.12 Information Collected Under Copyright Licenses

Some clinical trial data may be collected utilizing tools, questionnaires, or data dictionaries created by third parties from which a Data Provider has obtained a license. To maximize data utility, as much data (i.e., data variables) should be shared as is permitted by all applicable licenses (provided there are no additional concerns from a data privacy perspective).

A similar situation can exist with coding dictionaries as used for adverse events (**Section 4.8 Adverse Events & Medical History**) and medications (**Section 4.9 Concomitant Medications**). Some licenses allow for "full" sharing of coded data, whereas others allow sharing only according to narrow criteria (e.g., in accordance with requirements based on licensing contracts).

Data Providers should familiarize themselves with the relevant terms and conditions of their licenses. Data Providers are responsible to assess applicable licenses and copyrights while ensuring compliance with relevant data privacy laws.

Where possible and permitted under the license terms, it is recommended to:

- Share as many variables/data as possible within the limits of any applicable license.
- Ensure that any shared data, in the form of medication codes and adverse event codes, are covered by the Data Provider's specific license (**Section 4.8 Adverse Events & Medical History** and **Section 4.9 Concomitant Medications**)
- Include the license version and issue number in supporting documentation.
- Include any license limitations that may be relevant to secondary research teams in the supporting documents.
- Include the version and issue number of any questionnaires used in supporting documentation.

5.0 CONSIDERATIONS OF NOVEL AREAS FOR PRIVACY MEASURES

Scientific research is constantly evolving, introducing new insights and innovations. As advanced technologies become available, new types of data generation become possible along with new ways of analyzing data. Such novel developments must be considered when de-identifying/anonymizing data for research purposes. Data Providers sharing clinical data are encouraged to take novel developments into account and, if relevant, describe them in the supporting documentation. Below, are a few examples of novel areas for privacy measures to consider. However, given the complexity of these topics, they warrant a more detailed discussion that is beyond the scope of this paper. For that same reason, this section does not include any recommendation or compatible approach.

5.1 Data Derived from Genomic Data

While the information collected from clinical trial participants offers potentially great scientific utility, certain types of genetic/genomic data may represent high re-identification risk and, as such, must be removed.[viii]

Laws and regulations define omics and genetic data differently, and may encompass information relating to genes, gene variants, gene products (RNA or protein that results from the expression of the gene), and phenotypes.

It is preferable to retain information related to categorization of results from genotyping (e.g., wild type versus mutant) and chromosomal changes (e.g., Philadelphia chromosome positive [Ph+] or negative [Ph-]). However, from a data privacy perspective, it is encouraged (to the extent possible) to remove information pertaining to mutation analysis, single nucleotide polymorphisms (SNPs), and whole-genome and whole-exome sequencing. In case this data is to be included in further analyses, a genetics/omics expert should be consulted to determine the appropriate level of safeguards in a qualitative risk assessment.

5.2 Seasonality

Seasonality information can be an important feature to help understand disease patterns in certain therapeutic areas such as asthma or chronic obstructive pulmonary disease. In a therapeutic area where seasonality is relevant and was a component of the primary analysis, a month normalized to hemisphere can support many seasonality-related analyses. This means that months in the Southern hemisphere are mapped to their equivalents in the Northern hemisphere; for example, January, February, and March are mapped to July, August, and September as they both represent the winter season in their respective hemispheres. To derive such seasonality information at the participant-level, the original month (not day or year) must be maintained alongside continent (or hemisphere) in the dataset.

The aforementioned recommended approach to anonymize dates (offsetting, **Section 4.2 Dates**) and geographic information (generalized to continent, **Section 4.10 Geographic Location**) means that the information required to derive seasonality would not be present in all standard anonymized datasets. To maintain original month and hemisphere in the data, Data Providers would need to assess and adjust the anonymization approach for other variables accordingly to manage re-identification risk

[viii] DataCelerate® does not currently contain the type of genetics/genomic information that would identify individual patients either because it is not contained in the dataset or because it is removed by the contributor prior to upload.

(e.g., age bands may need to be larger). Organizations will need to determine: 1) whether seasonality is important for a given data, and 2) how the original month and hemisphere can be present in the dataset without increasing the risk of re-identification.

It is not recommended to include seasonality as an additional categorical variable (e.g., winter, spring) because this may not provide the specificity needed for seasonality analyses. If the recommended approach is followed for other variables, a categorical seasonality variable would pose a data quality challenge because it may be incompatible with the combination of anonymized geographic and date information for a given participant. In addition, categorical seasonality information also has the potential to reveal information about what has been altered through de-identification/anonymization (e.g., it may be possible to approximate the date offset interval to the nearest month, particularly if the date offset is not random).

Sharing of seasonality-related information requires further exploration to understand the number of datasets where seasonal information is relevant, and the impact of different de-identification/anonymization approaches for such datasets to accommodate original month and hemisphere.

6.0 SUPPORTING DOCUMENTATION

Pharmaceutical companies adhere to industry standards when processing clinical data (e.g., utilizing CDISC data formats, coding according to medical dictionaries). However, there remains high variability in data processing across companies, in part due to differences in anonymization technique. Some of the variability, together with a lack of transparency of the data modifications performed, leads to difficulties in interpreting the clinical data without having associated meta information at hand.

Providing supporting documentation is essential to the proposed methodology described in this paper. To enable greater transparency and increase scientific data utility, a key component of data sharing is to provide associated documentation that describes the context to the data structure and the modifications/derivations performed (e.g., if rare adverse events have been removed). This additional information will enable the potential Data User to assess a study dataset's suitability for inclusion in a particular secondary analysis and thereby ensure the scientific integrity of the subsequent results.

DataCelerate® provides a checklist of the supporting documents currently required as part of a study package upload[ix].

To help Data Providers determine which supporting documentation is useful to include, we have provided a Transparency Checklist in **Appendix 2** that outlines

[ix] It is recommended that any data sharing even outside DataCelerate® provides such contextual information.

the information we recommend be included as part of a study package. If an organization already has an anonymization report, containing the similar information as the Transparency Checklist, this report can be provided instead.

7.0 CLOSING REMARKS

Anonymization of clinical data is an evolving area, in both regulations and practice. As clinical data sharing and reuse matures and becomes more common in the pharmaceutical industry due to advances in data analysis technology, data privacy methodologies (including the one described in this paper) must be revised to allow users to protect the privacy of research participants and facilitate scientific data utility.

APPENDIX 1: GLOSSARY

The following table provides common acronyms and definitions relevant to key terms within this document. For additional related terms, please refer to the GDPR Data Reuse Initiative Glossary [16].

TERM	DEFINITION
CDISC	Clinical Data Interchange Standards Consortium
Data Modification	De-identifying, pseudonymizing, or anonymizing data using privacy preserving techniques and approaches, such as redaction or removal, to reduce the risk of re-identifying research participants from the clinical data. *Synonym: data transformation, data protection*
Data Provider	An individual or organization that is responsible for preparing and sharing data for reuse (e.g., sponsor). *Synonym: Data Contributors*
Data User	An individual or organization responsible for downloading and applying the shared data (e.g., researcher, company, data end user). *Synonym: Data Consumer*
Data Variable	CDISC Data Domain
Direct Identifiers	Data values that, on their own make it possible to uniquely identify a study participant. Direct identifiers could include Unique Subject ID (USUBJID), serial number or outliers.
Indirect Identifiers (also known as quasi-identifiers)	Data values that, in combination with others in the dataset make it possible to uniquely identify a study participant. Indirect identifiers could include demographic information, site identifier, vendor identifier, batch/lot numbers or study administrative data variables.
Non-identifiers	Another category which may exist within a clinical dataset. Non-identifiers are data variables or values that are either not linked to an individual trial participant or do not contribute to participant re-identification, examples include trial code/study ID, EUDRA or IND number, NCT number.
Research Participants	A voluntary, consenting individual participating in a clinical trial or research study.

	Synonyms: Study Participant, Trial Participant
MedDRA	Medical Dictionary for Regulatory Activities
SDTM	Study Data Tabulation Model
Sensitive Information	Sensitive Information is a subset of Personal Data that receives special protection because misuse could result in financial loss, identity theft, fraud, embarrassment, or significant reputational harm. Included in the category of Sensitive Information are: Data that could be used to discriminate, such as, but not limited to, data identifying one's race, ethnicity, religious or political beliefs, trade union memberships, physical or mental health, sexual orientation, or sex life, genetic or biometric data [17]
WHODRUG	WHO Drug Dictionary; WHO-DD

APPENDIX 2: EXAMPLE OF A COMPLETED TRANSPARENCY CHECKLIST

The Transparency Checklist should be included with each study package shared, to provide transparency to the data protection methods applied. Please refer to the Transparency Checklist on TransCelerate's Data Privacy Methodology website for a clean version.

Note: if an anonymization report has already been prepared, and it covers the elements described in the below Transparency Checklist, then sharing the anonymization report rather than filling in the Transparency Checklist is acceptable.

The following is an example of how the Transparency Checklist could look.

TRANSPARENCY CHECKLIST		
PART 1. Privacy Approach Applied		
1a. [MANDATORY] Specify the approach applied for each type of variable.	**Select one for each type of variable:** • **TransCelerate's Recommended Approach**	**Please elaborate on the approach applied (e.g., whether variables were removed or altered, rationale for why certain**

TRANSPARENCY CHECKLIST		
	• **TransCelerate's Compatible Approach** • **Other Approach**	**variables were removed/altered**
Unique Identifiers *NOTE: Where identifiers have been removed, the rationale should be described*	Recommended approach	Relevant unique identifiers have been scrambled
Dates	Recommended approach	Dates offset
Verbatim/Free Text	Recommended approach	All fields have been marked as redacted
Banding of Variables	Compatible approach	Semifixed banding applied to Height
Patient Demographics (sex, race, ethnicity)	Recommended approach	Patient ethnicity removed due to the low frequency in the dataset of some groups and there is no statistical evidence that indicates different ethnicities have an impact on the findings of the study
Data With Low Frequencies	Recommended approach	A minimal frequency of 1 out of 11 has been used for race
Sensitive Information	Recommended approach	All relevant values have been marked as redacted
Adverse Events & Medical History *NOTE: If any MedDRA levels are removed, please describe the reasons behind the removal.*	Compatible approach	LL has been removed due to limited trial size and rarity of disease. MedDRA version 27
Concomitant Medications	Recommended approach	no changes, version included in dataset

TRANSPARENCY CHECKLIST		
NOTE: The version information provided here unless it is provided in a variable in the dataset.		
Geographic Location	Recommended approach	Generalisation to "Middle East" region due to low number of participants and countries in region, other left at country level
Records of Participants Who Have Died	Recommended approach	N/A
1b. [MANDATORY] If appliable, please elaborate on the approach applied to the following variables.	x	
Information Collected Under Copyright Licenses	N/A	
Data Derived from Genomic Data	N/A	
Seasonality	N/A	
PART 2. Data Participants in Dataset		
2a. [MANDATORY] Has any individual participant's data been removed from the dataset due to anonymization requirements? Indicate Yes/No	No	
2b. [MANDATORY] If yes in 2a, provide the current number of participants included in the dataset.	N/A	
2c. [OPTIONAL] If yes in 2a, provide the rationale for removal.	N/A	
PART 3. Other Information		
3. [OPTIONAL] Indicate if your anonymization report has	No	

TRANSPARENCY CHECKLIST	
been provided in the study upload package or if one can be publicly accessed. If available, it is strongly recommended that you share your anonymization report or equivalent document. Please redact any information identifying the vendor before sharing. *If the anonymization report covers any of the other elements in this Transparency Checklist, there is no need to duplicate the information.*	
4. [OPTIONAL] Please describe the data format of the provided dataset, e.g., SDTM, ADaM and/or Other. *In DataCelerate®, indicate the data format specifications in the Transparency Checklist.*	SDTM
5. [OPTIONAL] Please indicate if the study is for an indication where seasonality is a key factor, any adaptions that were required to the methodology (e.g., how variables related to regions and dates have been protected).	No
6. [OPTIONAL] Please provide any other information that is considered helpful to a future research team.	13 participant records have been removed due to national EC requirements

APPENDIX 3: REFERENCES

In addition to TransCelerate member company input, the following external references contributed to the development of this document.

[1] Mello MM, Lieou V, Goodman SN. Clinical trial participants' views of the risks and benefits of data sharing. *N Engl J Med*. 2018;378(23):2202-2211. doi:10.1056/NEJMsa1713258.

[2] ISO/IEC 27559:2022(E) Information security, cybersecurity, and privacy protection – Privacy enhancing data de-identification framework. First edition 2022-11.

[3] TransCelerate Biopharma Inc. De-identification and Anonymization of Individual Patient Data in Clinical Studies: A Model Approach. 2016. Accessed October 20, 2022. http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-V2.0.pdf

[4] TransCelerate Biopharma Inc. A Privacy Framework for Clinical Data Reuse: Secondary Data Use in the Pharmaceutical Industry. October 2020. Accessed October 20, 2022. https://www.transceleratebiopharmainc.com/wp-content/uploads/2020/10/TransCelerate_A-Privacy-Framework-for-Clinical-Data-Reuse_Interpretations-of-Guidances-and-Regulations-Clinical_October-2020.pdf

[5] TransCelerate Biopharma Inc. Privacy Methodology for Data Sharing Solutions: Educational Toolkit for Clinical Data Reuse. Accessed October 20, 2022. https://www.transceleratebiopharmainc.com/assets/privacy-methodology-for-data-sharing-solutions/#undefined

[6] PHUSE Working Group. Data Transparency. Accessed October 20, 2022. https://advance.phuse.global/display/WEL/Data+Transparency

[7] TransCelerate BioPharma Inc. Welcome to TransCelerate. Accessed October 20, 2022. www.transceleratebiopharmainc.com

[8] European Union Agency for Cybersecurity. Data Pseudonymisation: Advanced Techniques and Use Cases. January 28, 2021. Accessed October 20, 2022. https://www.enisa.europa.eu/publications/data-pseudonymisation-advanced-techniques-and-use-cases

[9] European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use, v 1.3. September 22, 2017. Accessed October 20, 2022. https://www.transceleratebiopharmainc.com/assets/privacy-methodology-for-data-sharing-solutions/#undefined

[10] PHUSE Working Group. De-Identification Standard for SDTM 3.2 Version 1.0. May 20, 2015. Accessed October 20, 2022. https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+Standard+for+SDTM+3.2+Version+1.0.xls

[11] US Food and Drug Administration. Guidance Document: Collection of Race and Ethnicity Data in Clinical Trials. October 2016. Accessed October 20, 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/collection-race-and-ethnicity-data-clinical-trials

[12] World Health Organization. Anatomical Therapeutic Chemical (ATC) Classification. Accessed October 20, 2022. https://www.who.int/tools/atc-ddd-toolkit/atc-classification

[13] United Nations Department of Economic and Social Affairs Statistics Division. Methodology: M49 Standard. 2022. Accessed October 20, 2022. https://unstats.un.org/unsd/methodology/m49/

[14] General Data Protection Regulation (GDPR). Accessed October 20, 2022. https://gdpr-info.eu/

[15] UK *Data Protection Act 2018.* Accessed October 20, 2022. https://www.legislation.gov.uk/ukpga/2018/12/section/1

[16] TransCelerate Biopharma Inc. TransCelerate Interpretation of Clinical Guidances & Regulations: GDPR Data Reuse Initiative Glossary. October 2020. Accessed October 20, 2022. https://www.transceleratebiopharmainc.com/wp-content/uploads/2020/10/TransCelerate_Interpretations-Of-Clinical-Guidances-And-Regulations_GDPR-Glossary_October-2020.docx.

[17] General Data Protection Regulation (GDPR) Article 9. Processing of special categories of personal data. Accessed July 24, 2023. https://gdpr-text.com/read/article-9/