Clinical Data Transparency Initiative

DE-IDENTIFICATION AND ANONYMIZATION OF INDIVIDUAL PATIENT DATA IN CLINICAL STUDIES

A MODEL APPROACH

TransCelerate BIOPHARMA INC.

ACCELERATING THE DEVELOPMENT OF NEW MEDICINES

CONTENTS

BACKGROUND

TransCelerate BioPharma Inc., is a non-profit organization of biopharmaceutical companies focused on advancing innovation in research and development (R&D), identifying and solving common R&D challenges, increasing the quality of clinical studies and delivering more high-quality medicines to patients. Accordingly, TransCelerate has undertaken a commitment to enhance public health and medical and scientific knowledge and streamline regulatory compliance by facilitating the sharing and transparency of clinical trial information.

INTRODUCTION

The primary focus of this paper is to consider how de-identification and anonymization[a] techniques can be applied to individual patient data (IPD) in order to fulfil regulatory requirements relating to transparency and disclosure and to respond to research requests while simultaneously safeguarding the privacy of individuals (eg, participants and staff involved in conducting and evaluating clinical studies) for sharing in a non-public environment. This paper proposes which techniques to apply in order to conform to existing directives and regulatory guidance while balancing the utility of the de-identified data to the researcher. The TransCelerate "Clinical Study Reports Approach to Protection of Personal Data" paper[1] provides guidance on protecting study participant privacy in clinical study reports.

National data privacy laws,[2, 3, 4, 5] regulatory agency directives,[6, 7, 8] and several other guidances and papers address data protection and/or the sharing of individuals' personal data in the context of clinical research activities.[b] A common theme amongst these documents is "anonymized data" – that is, data from which all personal identifiers have been removed, such that it is no longer possible by all means reasonably likely to be used, to link the data back to an identifiable individual (including by combining that data with other data). Once data has been effectively and irreversibly anonymized, it will no longer fall within the scope of data protection and privacy laws. Note however that this concept is interpreted very narrowly by some regulators, who take the view that any "singling out" (i.e. any ability to distinguish one individual from another, even if it is not possible to tell who they are) triggers the application of data protection laws[16, 17]

One particular policy, European Medicines Agency Policy 0070[7, 8] states the importance of "...balancing the protection of study participant's privacy whilst retaining scientific value of the data". The EMA has undertaken a step-wise implementation of the policy. In the first phase, in force from 1st January 2015, only clinical reports, excluding IPD, will be published. In the second phase, various aspects relating to IPD will be clarified with stakeholders, addressing issues such as the submission of IPD for subsequent scientific review, and conditions and methods for providing access to IPD.

There are a number of techniques that can be used by data providers (eg, sponsor companies) to de-identify datasets prior to sharing. In order to provide increased benefit to the broader research community, the TransCelerate members aim to gain alignment across their member companies and with other industry groups' data de-identification and anonymization models and transparency principles [International Pharmaceutical Privacy Consortium (IPPC),[18] Pharmaceutical Research and Manufacturers of America/European Federation of Pharmaceutical Industries and Associations (PhRMA/EFPIA),[19] European Federation of Statisticians in the Pharmaceutical Industry (EFSPI), and Pharmaceutical User Software Exchange (PhUSE)]. This paper is intended to provide guidance aligned to regulatory policy, thus creating a common approach for the industry that can support activities in this area including meeting PhRMA/EFPIA commitments.[19]

a The terms 'de-identification' and 'anonymization' as used in this paper are defined under Key Topics below.

b Nothing in this paper should be construed as legal advice, nor does anything in this paper imply or warrant that use of this approach complies with applicable laws or regulations. Users implement the approach outlined in this paper at their own risk and bear the sole responsibility for ensuring their compliance with applicable laws and regulations in their respective jurisdictions.

ASSUMPTIONS AND CONSIDERATIONS

In defining the scope of this paper, the following assumptions were made:

- There is a legally binding data-sharing agreement between the data provider and the researcher prior to sharing data. This agreement should include clauses to impose data security principles to protect the confidentiality, integrity and availability of the data.
- The data provider has defined a secure method for sharing de-identified or anonymized data, ie, de-identified or anonymized data are shared in a controlled manner such as a password-protected environment

Furthermore, it is noted that case-by-case assessments may be required by data owners to determine the appropriateness of disclosing study information or particular datasets in special circumstances such as rare diseases, small populations, single-center trials, and low-frequency events.

DEFINING PROTECTED INFORMATION

This paper will use the term personally identifiable information (PII)[20] to describe protected information. The definitions of PII as well as of the protection of personal data[5] are provided in Appendix 1: Defining Protected Information.

SCOPE

The scope of this paper is defined as:

- The proposal of a model approach to de-identify and then anonymize personal information, providing an explanation of techniques that can be applied to structured raw (eg, SDTM) and/or reporting/analysis (eg, ADaM) datasets.
- All structured datasets are considered in scope of this paper although it is acknowledged that some, such as those containing genetic information, may require particular attention. Data providers may choose to exclude such datasets from their data-sharing agreements.

Specific systems, required technologies, and company confidential information contained within datasets are considered out of scope of this paper.

Sharing of data outside of the assumptions above, eg, public sharing, are out of scope of this version of the paper but will be addressed in a future release or separate document.

KEY TOPICS

For the purposes of this paper, we have defined de-identification and anonymization as follows, although we acknowledge the definitions for these terms may differ across other guidances (ie, industry, regulatory authorities), geographies and contexts.

» As noted in the PhUSE De-identification Standard for SDTM 3.2, "Data ***de-identification*** (verb: to de-identify) is the process by which a dataset is derived in such a way that the data subject is no longer identifiable"[21] ensuring that the risk of re-identifying a participant in a clinical trial, by all reasonably likely means to be used, is small. Various techniques are applied to de-identifying direct as well as indirect/quasi-identifiers since, when combined together and/or with other data could, increase the risk of re-identification. These techniques include: removing or re-coding identifiers, generalization of continuous data into categorical form, removing or redacting free text verbatim terms, and removing explicit references to dates. Participants' identification code numbers are de-identified by replacing the original code number with a new randomly generated code number.

» ***Anonymization*** is defined in this paper as a step subsequent to de-identification that involves irreversibly destroying all links between the de-identified datasets and the original datasets. This includes destroying the key code that was used to generate the new identification code number from the original, and destroying the deltas if dates were de-identified using the offset method.[22] Both the current EU Data Protection Directive[23] and the General Data Protection Regulation provide definitions of anonymization.[17] As noted by the Article 29 Working Party in their 2014 publication on Anonymization Techniques, "Recital 26 signifies that to anonymize any data, the data must be stripped of sufficient elements such that the data subject can no longer be identified. More precisely, the data must be processed in such a way that it can no longer be used to identify a natural person by using 'all the means likely reasonably to be used' by either the controller or a third party. An important factor is that the processing must be irreversible."[9]

Data providers should ensure plans for de-identification/anonymization of IPD are acceptable under relevant regional/national legislation.

In the United States, the HIPAA Privacy Rule de-identification standard published by the Department of Health and Human Services at 45 CFR 164.514 provides two alternative approaches to de-identification. These approaches (described below), often referred to as "Safe Harbor" and "Expert Determination", start with a shared principle of identifying direct study participant identifiers (eg, names, ID number) and indirect/quasi identifiers (eg, site code number, verbatim text, date of birth, and date of death), and then applying de-identification techniques. The definitions for these types of identifiers are described further in section De-Identification Steps.

It should be noted that the Safe Harbor approach is only defendable to some extent in the US and also assumes that the data controller has "no actual knowledge residual information can identify individuals" while Expert Determination is recognized across jurisdictions.

The HIPAA Privacy Rule: Safe Harbor Method

This first method under the HIPAA Privacy Rule[3] describes 18 types of identifiers that must be removed in order for the resultant datasets to be considered de-identified. The identifiers most commonly collected in clinical studies are: names (eg, investigators and vendors), contact numbers and addresses (eg, investigators' and vendors' telephone and fax numbers, and postal and email addresses), dates, device identifiers and serial numbers, photographic (or other comparable) images, characteristics (eg, verbatim text including reported adverse events, medical history, concomitant medications, and other comments), and any other unique identifying number (eg, treatment kit numbers) or code except a random identifier code.

Even after removing the 18 Safe Harbor identifiers, the resulting clinical data may nonetheless have a risk of re-identification of the data subject. In other words, the HIPAA Safe Harbor method may not be enough to satisfy the obligations to anonymize data under the EU Data Protection Directive or similar regulatory directives. Therefore, sponsors need to identify and remove any other PII that may still be present in the dataset ie, an **enhanced** Safe Harbor method is required.

The HIPAA Privacy Rule: Expert Determination Method

The second de-identification model under the HIPAA Privacy Rule involves, "A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:

i. Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and

ii. Documents the methods and results of the analysis that justify such determination."

The Guidance on De-Identification of Protected Health Information[14] provides a discussion of both methods, together with examples.

The Institute of Medicine issued their report on responsible sharing of clinical trial data in January 2015, "Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk".[24] Appendix B of the report includes a commissioned paper on "Concepts and Methods for De-identifying Clinical Trial Data". This appendix provides a general introduction to de-identification of data, how to measure and manage risk of re-identification, assessing the impact on data quality and governance and setting appropriate thresholds. While it covers a general overview of the topic, it focuses primarily on the quantification of risk of study participant re-identification and the HIPAA Expert Determination method, as outlined above.

It is the responsibility of each sponsor to define and measure for itself what the residual risk is and define an acceptable risk threshold in accordance with applicable laws and legislation.

The BMC Medical Research Methodology 2016 paper, "Protecting patient privacy when sharing patient-level data from clinical trials"[25] provides further discussion of best practices for de-identifying clinical trial data in secured environments. Even in the context of providing data in a secure, controlled-access model where data requests are reviewed and subject to data-sharing agreements, data providers may decide that a statistical assessment of the risk of re-identification (a key part of the Privacy Rule's Expert Determination approach) is necessary. As described in the PhUSE De-Identification Working Group paper "Providing De-Identification Standards to CDISC Data Models",[26] there may be residual re-identification risk under certain conditions, such as:

- The trial is for a rare disease;
- There are extreme values in the dataset;
- There are observable or knowable serious adverse events in the trial (eg, deaths and suicides);
- The dataset has extensive demographic and socio-economic information about the participants;
- The dataset includes detailed medical histories of the participants;

As the scope of clinical data shared in controlled environments increases, the risk of re-identification becomes more salient. While the data sharing agreement may reduce the probability of an attempt to re-identify, it may not manage all types of risks. Therefore, further data de-identification approaches may be needed.

DE-IDENTIFICATION STEPS

Once a particular datatype is singled out for de-identification, there are a number of approaches that can be used to de-identify participant-level data including generalization and randomization techniques.[9, 16]

Approaches are proposed in the following sections in line with recommendations made by PhUSE.[21] These include:

- Generalization techniques such as aggregation which can be used to organize continuous age data into age categories or to group countries into regional or continental level
- Suppression or masking where identifier values such as verbatim terms, names, and addresses are either set to blank or removed from the de-identified dataset
- A randomization approach, noise addition or data offsetting, as one method to de-identify

These approaches are summarized in Appendix 2: Summary of Approach to De-identification and an example of a dataset reflecting these techniques is provided in Appendix 3: A non-real Example Illustrating Removal of Personally Identifiable Information.

Direct Identifiers

As defined in ISO/TS 25237:2008,[27] direct identifiers are "data that directly identifies a single individual." It is further noted, "Direct identifiers are those data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain."

For any de-identification approach to be successful, it is necessary to first identify the datatypes that fit the category of "direct identifiers", and then apply a de-identification technique to those datatypes.

The PhUSE definition for direct identifiers refers to the possibility to re-identify a study participant alone or in combination with other direct identifiers (eg, address and name): One or more direct identifiers can be used to uniquely identify an individual (eg, subject ID, social security number, telephone number, exact address, etc.) It is compulsory to remove or de-identify any direct identifier.[21]

In order to protect participant privacy, the following direct identifiers should be modified ("re-coded"), and to anonymize the data, the key code that was used to generate these new random identifiers should be irreversibly destroyed.

- The original participant identifiers are replaced by new randomly generated identifiers (or code numbers). The dataset with the new random participant identifiers should be sorted by this new identifier so that the original participant order is changed.
- If the investigator/site identifier(s) are retained in the IPD, the investigator/site identifier(s) should be re-coded using new randomly generated identifiers (or code numbers). In order to maintain the relationship between participants and investigators/sites, all participants from one investigator/site should be assigned the same random investigator/site identifier in the de-identified dataset(s). If the investigator/site ID is used as part of another identifying variable (eg, unique subject identifier), it should be assigned the same value as the new random identifier.
- All other types of identifiers (eg, treatment kit numbers, device numbers, laboratory IDs) should be handled the same way, ie, be re-coded using new, randomly generated identifiers or removed/set to blank.
- When associated with a participant, investigator/site name and contact information, as well as that for any third-party vendors (such as laboratories and providers of imaging and biomarker data) should be set to blank or removed.

» An additional level of de-identification may be considered to further protect participant data privacy, eg, for studies with small numbers of participants within countries, investigators or sites. Identifier values could be set to blank, set to "--redacted--" or be aggregated into more general identifiers (eg, grouping of centers and/or countries by algorithms used in the original study analyses). Data providers would need to consider multiple data driven factors such as the number of participants and sites in each country, the size of the patient population, the disease state, and the impact on any future potential research analyses.

In order to maintain the relationship between data in connected studies, the same new identifiers (or code numbers) should be used across all datasets applicable to that single study. This includes both raw (eg, SDTM) and reporting/analysis (eg, ADaM) datasets.

In order to maintain the relationship between data in connected studies, such as a main study and its extension, the same new identifiers (or code numbers) should be used across all datasets applicable to both studies. This also applies to long-term follow-up studies where separate reports are published. This can be achieved by performing or repeating the data de-identification process for the initial study data at the same time as the extension/follow-up data. If the main study data have already been anonymized (ie, the key code destroyed) and provided to researchers, eg, directly after the main study is completed, a new set of randomly generated identifiers across all studies will be needed when the process is re-run after the extension study is also completed.

Indirect/Quasi Identifiers

As defined in ISO/TS 25237:2008,[27] indirect (or quasi identifiers) are "data that can identify a single person only when used together with other indirectly identifying data. It is further noted that "indirect identifiers can reduce the population to which the person belongs, possibly down to one if used in combination; eg, postcode, sex, age, date of birth."

The following sections describe techniques that can be applied to de-identify these variables.

Dates

Many different dates may be recorded in clinical studies, including visit dates, dates of birth, dates of adverse events, etc. Removal of the following date elements from datasets is required in order to achieve the Safe Harbor method of de-identification, "all elements of dates (except year) for dates directly related to the individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older".

There are two methods commonly used to de-identify dates, an "Offset Date" method and a "Relative Study Day" method. Both approaches require complete dates, therefore, data providers must apply imputation algorithms to any partial dates in order to use these methods.

Note: In some studies such as seasonal allergy or influenza studies, the actual calendar dates may be considered important to a researcher, however, in order to comply with the Safe Harbor de-identification method, even in these cases, actual dates still need to be de-identified. In some instances, it may be possible to use a generalization technique and provide the date in a categorical format. The definition of the categorical format will depend on a number of factors, such as the disease area, the seasonal definition required for the research proposal, and the countries in which the study was conducted. For an influenza study, the categorical format might be "peak flu season" and "out of flu season". The categorization would have to be done prior to de-identification and the "Relative Study Day" method should be used to de-identify the dates in this instance.

Offset Date Method

All dates are replaced with a new date generated using a random offset for each participant, and this offset is applied to all dates in the study for that participant. By using one offset for all dates for a participant, the relative distance between a participant's dates is maintained from the original dates to the de-identified dates.

This method could be implemented by having only one random offset for an entire study, and this would maintain the relative distance between dates recorded for different participants in that study as well. However, a drawback of having only one random offset to de-identify all dates in a study is that it may be perceived as not being as secure as having a different random offset for each participant since if the offset is identified for one participant, it is identified for all participants in that study. For this reason, an algorithm that assigns different random offsets to each participant in a study is considered a stronger approach when using this method.

Relative Study Day Method

If a variable containing Relative Study Day is not already present in the data provider's datasets, it is calculated for each observation as days relative to a reference date, eg, date of study entry or date of randomization. The same algorithm is applied to all dates across the study in order to maintain the relationship between events for each participant (eg, their visit schedule). All date variables are then removed from or set to blank in the de-identified datasets.

Combining Offset Date and Relative Study Day Methods

The offset date and relative day methods do not need to be interpreted as mutually exclusive, and both can be applied together. For example, SDTM and ADaM contain many variables that define relative study day (eg, variables ending in 'DY').

Date of Birth and Age

In order to adhere to the requirements of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule using the Safe Harbor method, additional requirements are stipulated to protect the privacy of participants aged over 89 years by aggregating their ages into a single category rather than presenting their exact age. De-identified datasets must also not display any dates indicative of age >89 years, eg, year of disease diagnosis or year of initiating a prior medication.

Thus, a de-identified dataset could contain:

- A variable containing the exact age for any participants aged ≤89 years, that is set to blank for any participants aged >89 years and
- A variable presenting age category for each participant, displaying either "≤89" or ">89".
- The associated variable, "Date of Birth" should be set to blank or removed regardless of the participants' age. [Note: HIPAA allows for retention of the year of birth for participants aged ≤89 years, but data providers may find it simpler to set to blank/remove regardless of participant age.]

In order to provide additional safeguards to participant privacy, generalization techniques could be applied to the data as described in EU Article 29 Data Protection WP216[9]. Using a generalization technique, date of birth would be removed from datasets, and age would be provided in a categorical form only, eg, using bands of 5 years such as <25 years, 25–29 years, 30–34 years,..., 85–89 years, >89 years. If an approach along these lines was implemented, any ages >89 years should be provided within one category of ">89 years" in order to comply with Safe Harbor requirements.

Additional attention is also required for studies enrolling pediatric participants, ie, those aged <18 years. In such studies, the following age bands, which are adopted from EudraCT reporting requirements[28] are recommended to allow greater data utility than the use of arbitrary bands:

- Preterm newborn (gestational age <37 weeks);
- Newborn infants (gestational age ≥37 weeks – 27 days after birth);
- Infants and toddlers (28 days–23 months);
- Children (2–11 years); and
- Adolescents (12–17 years).

Furthermore, adjacent categories could be combined if required to ensure appropriate levels of de-identification.

Medical Dictionaries and Coding

The most common dictionaries currently used by data providers are MedDRA for adverse events and diseases, and WHO Drug Dictionary for medications, although some data providers use their own in-house dictionaries. Dictionaries are upgraded at regular intervals, and datasets can be up-versioned as needed. For legacy studies, different dictionaries may have been originally used to code the data, eg, WHO-ART or COSTART (for adverse reactions).

Many dictionaries are used under license and so both researchers and data providers should be aware of any required licensing agreements before requesting or sharing any coded dictionary terms.

It is proposed that, wherever possible, and where dictionary licenses allow:

- Data providers using MedDRA to code adverse events/diseases provide all 5 levels of coding, namely, system organ class, high-level group term, high-level term, preferred term, and lowest level term.
- Data providers using WHO Drug Dictionary to code medications provide trade names and ingredients Note: The researcher may need to sign additional agreements in order to include WHO Drug Dictionary-coded terms, which can only be shared under certain circumstances.

One caveat to this approach is the presence of lowest level terms and product names of low frequency. These may need further review and subsequent aggregation with respect to protecting participant data privacy.

Legacy data may not have been coded originally or may have been coded using earlier versions of dictionaries or completely different dictionaries than used for current data, eg, initially coded using COSTART and now using MedDRA. Data providers should provide any coded terms available in the datasets (as permitted by dictionary licenses), ideally including lowest level terms so that the researcher can code using whichever dictionary and version is most appropriate for their purposes.

When providing coded terms, data providers should also provide the name and version number of each dictionary used wherever possible.

Free-Text Verbatim Fields

Redaction of free-text verbatim fields is necessary because such fields may contain participant-specific information and therefore may allow identification of that participant. In general, all free-text verbatim terms and comments variables should be set to blank or removed if redaction is required for every record in a dataset.

Certain free-text fields (or parts thereof) may be considered for retention in their original form if removal of this information would impact the scientific value of the dataset, eg, a free-text field in an oncology study where tumor site was recorded. Such fields should be reviewed carefully in a record-by-record fashion to ensure they do not contain personal information. If personal information is found in a field, the field (or part thereof) of the affected records should be replaced with the value of "--redacted--" in order to show that the original value was redacted for the purpose of de-identification thus highlighting that it was not a null field in the original dataset.

Sensitive Information or Low-Frequency Events

Sensitive information is data which can be quite concerning or potentially socially or financially harmful to an individual when revealed publicly. Examples of sensitive information include studies with rare diseases (eg, small denominators where the total eligible patient population is small), rare events (eg, small numerators), genetic information, extreme or outlier values of common information (eg, height, weight, body mass index), and behaviors such as illicit drug use or "risky behavior."

Low-frequency data is data that occurs within the dataset in very small numbers. For example, if one out of 200 participants experiences migraines, this may be considered a low-frequency event. The threshold for what is low-frequency in each data set must be defined via a risk assessment. Low-frequency events may or may not also be sensitive information.

Sensitive data and low-frequency data may or may not overlap, but must be managed separately when approaching data anonymization. Sensitive data are not necessarily re-identifying while they would harm the patient in case of re-identification. If we can ensure that the data is not personal anymore, sensitive data could be kept. Each sponsor must decide how they want to approach sensitive data. The data provider will need to balance the extent of measures taken to address required participant privacy while maintaining data utility, since data such as rare adverse events may be the key information desired by the researcher to perform their analyses.

Common de-identification techniques for both sensitive and low-frequency data types include adding noise (eg, using an offset method for dates) or aggregating data (eg, defining age bands). It is recommended that data providers employ such techniques when it is considered that participant data privacy specifically requires it. Other data items may have an increased sensitivity, and therefore, additional steps may be required to further protect participant data privacy such as setting variables to blank or replacing the "sensitive" records (or parts thereof) as "--redacted--".

QUALITY CHECKS

It is important that data providers perform a validation and review of their anonymization process to ensure that all necessary data have been de-identified appropriately and consistently between the datasets. Given that the destruction of a key code is a uni-directional step (ie, cannot be reversed), this review would need to be performed prior to destroying the key code that links the de-identified datasets to the original datasets.

The enhanced Safe Harbor approach combines removal of the relevant 18 HIPAA identifiers with the removal of additional personal information that may be present in a study dataset. Automated approaches provide benefits in terms of standardization and efficiency, however, the approach taken will need to be configured so that all variables to be de-identified and all variables to be redacted are identified correctly. It is recommended that data providers do not rely on a single technique, eg, running a "de-identification macro", to define their whole process. A manual review of datasets is strongly encouraged to identify variables or records requiring de-identification.

PROCESS RECOMMENDATIONS

Data providers need to manage both requests for de-identified data and the controlled access to these data. One option for managing the de-identification process is to join multi-sponsor solutions to manage research requests such as the Yale Open Data Access (YODA) Project[29] or www.clinicalstudydatarequest.com[30]. Alternatively, data providers or they may define their own processes.

Review of Requests

Regardless of approach, data providers should ensure that there are processes in place to review research requests, assess feasibility and scientific value,[31] confirm the research team are qualified to conduct the proposed research (eg, qualified statistical resource), and approve valid requests. Some data providers may use an external and independent group to perform some of these steps.

Data-Sharing Agreements

A legally binding agreement is highly recommended before any data are shared by a data provider including clauses to not share the data with any third party and to prohibit the researcher from attempting to re-identify individuals from the de-identified data.

Data and Documents Provided

In order for researchers to be able to perform their analyses, they may require accompanying documentation such as the anonymized protocol and amendments, annotated case report form, anonymized statistical analysis plan, dataset specifications, and anonymized clinical study report. For legacy studies, some of this information may not be available. There should be clear communication from the data provider

to the researcher regarding what documentation will be provided along with the requested de-identified datasets. All documentation provided to the researcher should be reviewed carefully to ensure that it does not contain any privacy information prior to sharing. This topic is addressed in the TransCelerate "Protection of Personal Data in Clinical Documents – A Model Approach" paper.[32] It is also recommended that data providers document the de-identification and anonymization steps they applied, and provide details of these steps to researchers.

Data providers may choose to provide only the de-identified/anonymized datasets required by the researcher of an approved request, or they may choose to provide all such datasets of the study/studies requested.

Data Anonymization

When preparing the multiple datasets for a single recipient researcher, it is recommended data providers de-identify all datasets for a study simultaneously. By doing this, although requiring additional resource up-front, it allows future requests from the same study to be handled more efficiently. These efficiencies include reducing the time taken to provide the de-identified data to future researchers, reducing or eradicating duplication of effort, eg, re-running the de-identification process and storing multiple versions of the resultant datasets. It is also recommended that de-identified/anonymized data are stored by the data provider in a separate location from the original data.

Case-by-case assessment may be required by data providers when data have been collected in the local language(s) of the countries in which the study was performed. The data providers' processes may require translation of the data (eg, into English) before it can be de-identified.

Anonymising IPD can be used to support anonymising clinical documents (eg, clinical study report). Anonymising IPD first can increase the efficiency and consistency of anonymising clinical documents that are the basis of the content of the clinical documents.

Data Access

An important consideration for data providers is the technical process for disclosure of requested datasets – or put simply, how the data are shared. Given that data providers retain some accountability with regards to unauthorized access to data, the model approach described in this paper assumes that de-identified data be provided to researchers of approved requests with controls on access, in which the researchers can perform their analyses and from which they can download their results. Generally, this is accomplished by the data provider storing the datasets within its environment and simply granting the researcher a temporary access portal through which the researcher can access the requested datasets. Access to data may be via a multi-sponsor environment or a data provider-specific solution. Consideration should also be given to the length of time that this access is available and to whom the access is granted. It is strongly recommended that access is granted on a named researcher basis, and data should only be shared in a controlled manner such as a password-protected environment.

The extent to which study data are disclosed is at the discretion of the data provider and may differ for current and legacy data; eg, reporting/analysis datasets may not exist for legacy studies and so only raw data would be available. With a move to CDISC, datasets will be in a more standardized format allowing for the provision of SDTM and ADaM datasets. Further description of CDISC, SDTM, and ADaM standards can be found at www.cdisc.org.[26]

Regardless of the choice of request management system and data access system, standardization of the de-identification process across data providers is strongly encouraged to most readily facilitate research analyses.

Review of Procedures

Data providers should also review their processes on a regular basis. This is to ensure that their approach remains robust as more data from more sources become available in the public domain, and as more advanced tools are developed, ie, new external information cannot be used to infer participant identities within previously de-identified datasets.

CONCLUSION

Increasingly, data providers are defining algorithms for de-identification and anonymization of data. This paper focuses on applying an enhanced Safe Harbor approach, supplementing with expert determination methods as needed. By applying a common approach across data providers, the utility of the de-identified and anonymized datasets will be increased such that datasets from multiple data providers can be combined more easily to facilitate meta-analyses.

As technology advances and more data become accessible, methodology will require regular review to ensure that the balance between data utility and required participant data privacy is appropriately maintained. Methods for quantifying risk of participant re-identification may need to be employed more frequently, particularly as external data sources grow and data linkage techniques improve.

The changing regulatory landscape means that anonymization and sharing of IPD should no longer be considered as a standalone task. Strategies will need to be developed which can eventually integrate and increasingly automate processes for anonymization of both IPD and reports This will facilitate proactive sharing (eg, EMA Policy 0070[8]), bearing in mind the very different access context – controlled/ secure vs. public access. In fact, this integrated approach will be aided by better planning at trial concept stage. This is in line with the Institute of Medicine's[11] recommendation that every trial should have a "data-sharing plan", where all downstream data-sharing requirements can be pre-specified.

Data de-identification and anonymization of IPD in clinical studies is an evolving area. Data privacy legislation is changing[33], the number of external sources of data which could potentially be link to clinical data is increasing, and requirements for sharing IPD and information are also increasing. The need to integrate data-sharing activities into day-to-day clinical trial activities never been greater. Furthermore, there is a need to regularly evaluate the effectiveness of anonymization techniques given the advances in data linkage techniques.

REFERENCES

1. TransCelerate BioPharma Inc. CSR Redaction of Privacy Information - Clinical Study Reports Approach to Protection of Personal Data. Available at http://www.transceleratebiopharmainc.com/wp-content/uploads/2014/08/TransCelerate-CSR-Redaction-Approach.pdf. Published August 28, 2014. Accessed March 6, 2015.
2. Code of Federal Regulations - Title 45: Public Welfare, Subtitle A §164.514: Available at http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title45/45cfr164_main_02.tpl. Accessed April 6, 2015.
3. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Available at http://privacyruleandresearch.nih.gov/pr_08.asp#8a.
4. Council Regulation (EC) 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data [2008] OJ L193/7. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV%3Al24222. Accessed April 6, 2015.
5. Council Directive (EC) 95/46 of the European parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046. Accessed April 6, 2105.
6. Regulation (EU) 536/2014 of the European parliament and of the council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC [2014] OJ L158/1. Available at http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32014R0536. Accessed April 6, 2015.
7. European Medicines Agency: EMA/240810/2013 - Publication of clinical data for medicinal products for human use. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf, October 2, 2014.
8. European Medicines Agency: EMA/90915/2016 - External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2016/03/WC500202621.pdf, March 2, 2016.
9. Article 29 Data Protection Working Party: 0829/14/EN WP216 - Opinion 05/2014 on Anonymisation Techniques. Available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf. Published April 10, 2014. Accessed March 10, 2016.
10. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. BMJ. 2010;340:c181.
11. Institute of Medicine. Discussion Framework for Clinical Trial Data Sharing: Guiding Principles, Elements, and Activities. Available at http://www.nationalacademies.org/hmd/Reports/2014/Discussion-Framework-for-Clinical-Trial-Data-Sharing.aspx. Published January 14, 2014. Accessed March 6, 2015.
12. Information Commissioner's Office. Anonymisation: managing data protection risk code of practice. Available at https://ico.org.uk/media/1061/anonymisation-code.pdf. Published November 20, 2012. Accessed March 6, 2015.
13. Hughes S, Wells K, McSorley P, Freeman A. Preparing individual patient data from clinical trials for sharing: the GlaxoSmithKline approach. Pharm Stat. 2014;13(3):179-184.
14. Office for Civil Rights. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Available at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html. Published November 26, 2012. Accessed March 6, 2015.
15. Jack Shostak; Duke Clinical Research Institute. De-Identification of Clinical Trials Data Demystified. Available at www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf. Published 2006. Accessed March 6, 2015.
16. Article 29 Data Protection Working Party: 0829/14/EN WP216 - Opinion 05/2014 on Anonymisation Techniques. Available at http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf. Published April 10, 2014. See also Recital 26 of the EU General Data Protection Regulation 2016/679.
17. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679. Accessed November 20, 2016
18. International Pharmaceutical Privacy Consortium. White Paper on Anonymisation of Clinical Trial Data Sets. Available at http://pharmaprivacy.org/assets/activities/IPPC_White_Paper_Anonymisation_Clinical_Trials_Data.pdf. Published October 13, 2014. Accessed March 6, 2015.
19. PhRMA/EFPIA. Principles for Responsible Clinical Trial Data Sharing: Our Commitment to Patients and Researchers. Available at http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf. Published July 18, 2013. Accessed March 6, 2015.
20. Executive Office of the President Office of Management and Budget. Memorandum 07-16 - Safeguarding Against and Responding to the Breach of Personally Identifiable Information. Available at https://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf. Published May 22, 2007. Accessed March 6, 2015.

21. *PhUSE De-Identification Working Group paper "Providing De-Identification Standards to CDISC Data Models". Available at http://www.phusewiki.org/docs/Conference%202015%20DH%20Papers/DH01.pdf. Accessed March 11, 2016.*

22. *The Anonymization Decision-Making Framework. Available at http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf. Published 2016. Accessed December 8, 2016. Reference Section 2.1.3*

23. *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046. Accessed November 20, 2016*

24. *Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. 2015. Available at https://www.nap.edu/search/?term=Sharing+Clinical+Trial+Data%3A+Maximizing+Benefits%2C+Minimizing+Risk. Accessed March 11, 2016.*

25. *Protecting patient privacy when sharing patient-level data from clinical trials. BMC Medical Research Methodology 2016. Available at http://link.springer.com/article/10.1186/s12874-016-0169-4. Published July 8, 2016. Access September 7, 2016.*

26. *Clinical Data Interchange Standards Consortium. Available at http://www.cdisc.org. Accessed March 6, 2015.*

27. *"Pseudonymization" – new ISO specification supports privacy protection in health informatics. Available at http://www.iso.org/iso/home/news_index/news_archive/news.htm?refid=Ref1209. Published March 10, 2009. Accessed August 10, 2016.*

28. *European Clinical Trials Database (EudraCT) V10. Available at https://eudract.ema.europa.eu/index.html. Accessed August 10, 2016.*

29. *The Yale Open Data Access (YODA) Project. A New Approach to Data Access and Transparency. Available at http://medicine.yale.edu/core/projects/yodap. Accessed March 6, 2015.*

30. *ClinicalStudyDataRequest.com. Available at https://clinicalstudydatarequest.com. Accessed March 6, 2015.*

31. *U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Guidance for Industry - E6 Good Clinical Practice: Consolidated Guidance. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073122.pdf. Published April 1996. Accessed March 6, 2015*

32. *TransCelerate Biopharma Inc. Clinical Data Transparency initiatives. http://www.transceleratebiopharmainc.com/assets/clinical-data-transparency/. Accessed November 8, 2016*

33. *European Commission. Agreement on Commission's EU data protection reform will boost Digital Single Market. Available at http://europa.eu/rapid/press-release_IP-15-6321_en.htm. Published December 15, 2015. Accessed March 11, 2015.*

APPENDIX 1: DEFINING PROTECTED INFORMATION

Directive 95/46/EC[4] describes the protection of **Personal Data** where personal data is defined as "any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."

OMB 07-16[17] states that **Personally Identifiable Information (PII)**, "refers to information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."

APPENDIX 2:
SUMMARY OF APPROACH TO DE-IDENTIFICATION

Attribute	Approach to De-identification
Identification Number *(eg, participant, investigator, site, laboratory)*	Consistently replace these original identifiers with new, randomly generated identifiers or set to blank. When the instance arises that site number cannot be removed due to losing the subject ID uniqueness, all participants from one investigator/site should be assigned the same random investigator/site identifier in the de-identified dataset(s). It is recommended to blank out identifiers other than participant numbers in order to not jeopardize data privacy.
Names *(eg, participant, investigator, site, contractor, supplier, vendor, company staff)*	Set to blank or remove.
Contact Information *(eg, participant, investigator, site, vendor - postal address, phone, fax, email)*	Set to blank or remove.
Country	Retain as in the original dataset unless this is considered to jeopardize data privacy. If so, set to blank, set to '--redacted--', or aggregate (data driven decision).
Dates	Consistently apply either "Offset Date method" or "Relative Study Day" method.
Participant Date of Birth	Set to blank or remove.
Age	The de-identified dataset could contain: » Exact age if ≤89 years, and set to blank if >89 years » Sub-category options: – Age category as "≤89" or ">89", or – Age category as ">89" and categorize ≤89 into smaller age bands eg, <25, 25-29, 30-34,...,85-89 Note: Age category ">89" must not be sub-divided
Adverse Event/Medical History/ Concomitant Medication/Procedure	Set verbatim-level terms to blank or remove. If permitted by dictionary licenses, provide coded terms, eg, ideally at least the lowest-level term (LLT) for MedDRA-coded adverse events and diseases, and trade name and drug ingredients for medications coded using WHO Drug Dictionary.
Other free-text fields *(eg, comment fields)*	In general, set to blank or remove. If considered important for retention, set records (or parts thereof) containing privacy information to "--redacted--".
Other indirect identifiers *(eg, rare disease or rare adverse event, extreme values, unusual treatment)*	Retain unless considered to jeopardize participant data privacy. If so, set to blank, set to "--redacted--", or aggregate as needed.

APPENDIX 3:
A NON-REAL EXAMPLE ILLUSTRATING REMOVAL OF PERSONALLY IDENTIFIABLE INFORMATION

Center ID	Investigator ID	Investigator Name	Subject ID	Unique subject ID	Age (years)	AE StartDate	AE End Date	Verbatim Term
00123	279344	Dr Smith	5	TJF4392.0005	57	29DEC2010	27JAN2011	Headache
00123	279344	Dr Smith	2	TJF4392.0002	72	10JAN2011	06APR2011	Nausea
00123	279344	Dr Smith	1	TJF4392.0001	91	25MAR2011	12AUG2011	Cold
05678	333721	Dr Jones	19	TJF4392.0019	85	14OCT2010	20OCT2011	Cold
05678	333721	Dr Jones	4	TJF4392.0004	53	24MAY2011	.	Headache
05678	333721	Dr Jones	23	TJF4392.0023	76	01MAR2011	15MAR2011	Pain

NEW Center ID	***NEW*** Investigator ID	***REMOVE*** Investigator Name	***NEW*** Subject ID	***NEW*** Unique Subject ID	***REMOVE*** Ages >89	***CREATE NEW*** Age Category	***ADD*** Offset to each date	***ADD*** Offset to each date	***REMOVE*** Verbatim Term

Center ID	Investigator ID
03145	148227
03145	148227
03145	148227
90876	687208
90876	687208
90876	687208

Subject ID	Unique subject ID	Age (years)	Age Category (years)	AE StartDate	AE End Date
8754	TJF4392.8754	57	<=89	02FEB2011	03MAR2011
5681	TJF4392.5681	72	<=89	09NOV2010	03FEB2011
1475	TJF4392.1475	.	>89	03JUL2011	20NOV2011
1457	TJF4392.1457	85	<=89	06JUL2010	12JUL2011
2214	TJF4392.2214	53	<=89	03MAY2011	
2236	TJF4392.2236	76	<=89	08MAR2011	22MAR2011

Alternative approaches include: assigning the same new center ID to low-recruiting centers; removing investigator ID; replacing dates with Relative Study Day.