TransCelerate BIOPHARMA INC.
ACCELERATING THE DEVELOPMENT OF NEW MEDICINES

Data De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach

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, thus increasing the quality of clinical studies and delivering more high-quality medicines to patients. The biopharmaceutical members of TransCelerate are committed to enhancing public health and medical and scientific knowledge through the sharing and transparency of clinical trial information.

Introduction

The primary focus of this paper is to consider how de-identification and anonymization[1] techniques can be applied to individual patient data (IPD) in order to fulfil transparency, disclosure and research requests while safeguarding the privacy of individuals (eg, participants and company staff). 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.

National data privacy laws [1, 2, 3, 4], regulatory agency directives [5, 6] and several other guidances and papers [7, 8, 9, 10, 11, 12, 13] relate to data protection and/or the sharing of individuals' personal data. Data privacy laws consider that if personal data are removed or de-identified and subject code identifiers cannot be linked back to specific individuals, then the data are no longer considered to be personal data. EMA policy 0070 [6] discusses the importance of '…balancing the protection of patient's privacy whilst retaining scientific value of the *data'* and also references the concept of '*anonymisation'* versus '*pseudonymisation'*.

There are a number of techniques that can be used by data providers (eg, Sponsor companies) to adequately 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 (IPPC [14], PhRMA/EFPIA [15], EFSPI/PSI and PhUSE). This paper is intended to provide input into regulatory policy, thus creating a common approach for the industry that can support activities in this area including meeting PhRMA/EFPIA commitments [15].

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

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 prohibit the researcher from further sharing the de-identified/anonymized data, and from attempting to identify individuals from these 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 studies, and/or small populations.

Defining Protected Information

This paper will use the term Personally Identifiable Information (PII) [16] to describe protected information. The definitions of PII as well as of the protection of personal data [4] and protected health information 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 raw (eg, SDTM) and/or reporting/analysis (eg, ADaM) datasets.
- All datasets are considered in scope of this paper though 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 and required technologies are considered out of scope of this paper.

Key Topics

For the purposes of this paper we have defined de-identification and the subsequent step of anonymization as follows, although different definitions for these terms are used in other guidances and contexts.

- **De-identified** protected health information (PHI) is defined in the HIPAA Privacy Rule, Code of Federal Regulations, 45CFR164.514, as 'Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.' Thus de-identifying the data includes removing or recoding identifiers, 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 random code number.

- **Anonymization** is a step subsequent to de-identification that involves destroying all links between the de-identified datasets and the original datasets. The key code that was used to generate the new identification code number from the original is irreversibly destroyed (ie, destroying the link between the two code numbers).

Data providers should evaluate all relevant regional/national legislation.

The standard in 45 CFR 164.514 provides 2 alternative approaches to de-identification. These methods are described below:

Safe Harbor Method

This method [2] describes 18 types of identifiers that must be removed (eg, by deleting, recoding or redacting) 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, investigator and vendor 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.

The 'Safe Harbor' method (and the 18 types of identifiers) is not focused on clinical trial data and therefore data providers using a de-identification method based on this approach also remove and review other personal information that may be present in the dataset, **ie, use an enhanced method.**

Expert Determination Method

This 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 [12] provides a discussion of both methods, together with examples.

The Institute of Medicines issued their report on responsible sharing of clinical trial data in January 2015, "Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk" [9]. 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. Whilst it covers a general overview of the topic, it focuses primarily on the quantification of risk of patient re-identification and the HIPAA Expert Determination method, as outlined above. We consider that the assessment of risk is a decision to be

made on a data provider by data provider basis, and the Institute of Medicines report outlines methods for doing this.

Both the Safe Harbor and Expert Determination approaches start with a shared principle of identifying direct (e.g. ID number) and quasi (e.g. date of birth and date of death) identifiers, and applying de-identification techniques. 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 Expert Determination approach) is not necessary in most cases. For this reason, the approach outlined in this paper is based primarily on the enhanced Safe Harbor method. For certain datasets or situations (e.g. rare diseases/small populations) sponsors may choose to employ expert determination methods to supplement that approach.

De-identification Steps

Recommended approaches to de-identification of various types of data are proposed in the following sections. These are summarized in Appendix 2: Summary of Approach.

Recoding Identifiers

In order to protect participant privacy, the following 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.

- A new randomly generated identifier (or code number) is required for each participant. The dataset with the new random participant identifiers should be sorted by this new identifier so that the original participant order is changed.
- 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).
- The 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.
- 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.

In order to maintain the relationship between different records in all datasets of a study, 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, with the key code destroyed) and provided to researchers eg, directly after the main study completed, there will need to be a new set of randomized identifiers across all studies when the process is re-run after the extension study also completes.

Handling Dates

Many different dates may be recorded in clinical studies including visit dates, dates of birth, and 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 order to protect participants’ privacy.

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 their original dates to their 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 therefore 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.

Handling Date of Birth and Age

In order to adhere to the requirements of the 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 a prior medication was started.

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 [7]. Using a generalization technique, date of birth would be removed from datasets, and age would be provided in a categorical form only, for example, 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.

Medical Dictionaries and Coding

The most common dictionaries currently used by data providers are MedDRA for adverse events and diseases, and WHO Drug for medications, though 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 to code medications provide trade names and ingredients. Note: Data providers may need to sign additional agreements in order to include WHO Drug 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 (dictionary licenses allowing), 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 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 and Low Frequency Events

Examples where information may be particularly sensitive, and a data provider may choose not to share, or may choose to employ additional de-identification techniques in order to further protect participant privacy, include studies with rare diseases (eg, small denominators where the total eligible patient population is small), rare events (eg, small numerators), genetic information, extreme values (eg, height, weight, BMI), or sensitive data (eg, illicit drug use or "risky behavior").

Other data items may be of 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) with "--redacted--".

Alternative approaches 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.

The data provider will need to assess the balance between the extent of measures taken to address required participant privacy while maintaining data utility, since data such as rare adverse events may be the exact information required by the researcher to perform their analyses.

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 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 an "de-identification macro", to define their whole process, and that 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. They may join multi-sponsor solutions to manage research requests such as the Yale Open Data Access (YODA) Project [17] or www.clinicalstudydatarequest.com [18], 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, confirm qualified statistical resource[2] in the research team, 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 must exist 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 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 protocol and amendments, annotated case report form, statistical analysis plan, dataset specifications, and a 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 with the de-identified datasets. All documentation provided to the

[2] Confirmation of qualified statistical resource is in line with ICH E6, Good Clinical Practice [20].

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 'Clinical Study Reports Approach to Protection of Personal Data' paper [19]. 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

It is recommended to de-identify all datasets for a study at the same time. 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 to 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.

Data access

In order to provide additional safeguards against unauthorized data access, 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 researcher can perform their analyses and from which they can download their results. 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. For example, 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, SDTMs and ADaMs standards can be found at www.cdisc.org [21].

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 or 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.

References

1. 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)
2. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule - http://privacyruleandresearch.nih.gov/pr_08.asp#8a
3. 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
4. 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
5. 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
6. European Medicines Agency: EMA/240810/2013 - Publication of clinical data for medicinal products for human use. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf, October 2014
7. Article 29 Data Protection Working Party: 0829/14/EN WP216 - Opinion 05/2014 on Anonymisation Techniques. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, April 10, 2014
8. 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.
9. Institute of Medicine. Discussion Framework for Clinical Trial Data Sharing: Guiding Principles, Elements, and Activities. http://www.iom.edu/Reports/2014/Discussion-Framework-for-Clinical-Trial-Data-Sharing.aspx. Published January 14, 2015. Accessed March 6, 2015.
10. Information Commissioner's Office. Anonymisation: managing data protection risk code of practice. https://ico.org.uk/media/1061/anonymisation-code.pdf. Published November 20, 2012. Accessed March 6, 2015.
11. 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.
12. 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. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html. Published November 26, 2012. Accessed March 6, 2015.
13. Jack Shostak; Duke Clinical Research Institute. De-Identification of Clinical Trials Data Demystified. www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf. Published 2006. Accessed March 6, 2015.
14. International Pharmaceutical Privacy Consortium. White Paper on Anonymisation of Clinical Trial Data Sets. http://pharmaprivacy.org/assets/activities/IPPC_White_Paper_Anonymisation_Clinical_Trials_Data.pdf. Published October 13, 2014. Accessed March 6, 2015.

15. PhRMA/EFPIA. Principles for Responsible Clinical Trial Data Sharing: Our Commitment to Patients and Researchers. http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf. Published July 18, 2013. Accessed March 6, 2015.
16. Executive Office of the President Office of Management and Budget. Memorandum 07-16 - Safeguarding Against and Responding to the Breach of Personally Identifiable Information. https://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf. Published May 22, 2007. Accessed March 6, 2015.
17. The Yale Open Data Access (YODA) Project. A New Approach to Data Access and Transparency. http://medicine.yale.edu/core/projects/yodap. Accessed March 6, 2015.
18. ClinicalStudyDataRequest.com. https://clinicalstudydatarequest.com. Accessed March 6, 2015.
19. TransCelerate Biopharma Inc. CSR Redaction of Privacy Information - Clinical Study Reports Approach to Protection of Personal Data. http://www.transceleratebiopharmainc.com/wp-content/uploads/2014/08/TransCelerate-CSR-Redaction-Approach.pdf. Published August 28, 2014. Accessed March 6, 2015.
20. 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.
21. Clinical Data Interchange Standards Consortium. http://www.cdisc.org. Accessed March 6, 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 [16] 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.'

Protected Health Information (PHI) is defined in the 45 CFR 160.103 as, 'individually identifiable health information transmitted or maintained by a covered entity or its business associates in any form or medium. It includes demographic information related to health data collected from an individual that can reasonably be used to identify the individual. Additionally, PHI is information created or received by a health care provider, health plan, employer, or health care clearinghouse; and relates to the past, present, or future physical or mental health or condition of an individual.'

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. 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). For identifiers other than participant numbers: If considered to jeopardize data privacy, can set to blank, set to '--redacted--', or aggregate.
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	• Display exact age if <=89 years, and set to blank if >89 years • Display age category - either as '<=89' and '>89', or sub-categorizing '<=89' into multiple age bands eg, <25, 25-29, 30-34,…,85-89. (Age category '>89' must not be sub-divided).
Adverse Event / Medical History / Concomitant Medication / Procedure	Set verbatim level terms to blank or remove. Dictionary licenses allowing, provide coded terms eg, ideally at least the lowest level term (LLT) for MedDRA-coded adverse events and diseases; and trade name and

Attribute	Approach to De-identification
	ingredients for medications coded using WHO Drug.
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.

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.