PHUSE Good Transparency Practice

phuse.global

Contents

Revision History

Version	Date	Summary
1.0	15.08.25	

Introduction

Good Transparency Practice (GTP) is a guideline developed by the PHUSE Data Transparency Working Group to create a set of best practices to govern the anonymisation of clinical trial data, for external sharing or disclosure. Although there are transparency initiatives across the globe, with differing guidelines, the common goal is to uphold patient privacy and data utility to the highest standards. Data is a vital asset, and when shared, helps advance science and increase public confidence in clinical trial development. While guidance on the proper conduct of clinical studies is outlined by the International Council for Harmonisation (ICH) in Good Clinical Practice (GCP) [1], there are no such 'good practice' guidelines specifically for anonymisation and de-identification and subsequent sharing of data from these clinical studies. Anonymised data does not have the same restrictions as the original untransformed data, therefore making it more feasible to use for data sharing. The original data can only be for its primary use unless additional informed consent is gained from the data subjects. However, obtaining informed consent for future data sharing can be very difficult, as protections can vary between sites, countries, continents and regions. Alternatively, informed consent from data subjects is not always required for sharing anonymised data, which can allow for greater freedom in data sharing.

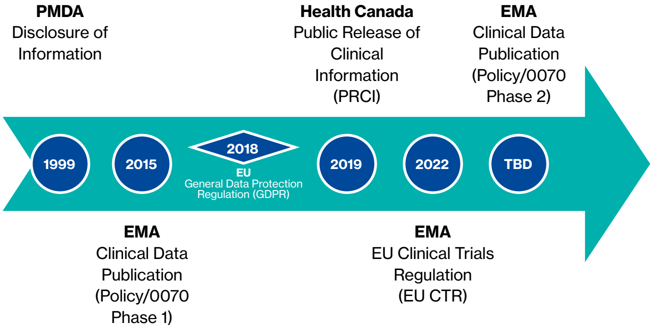

PMDA
Disclosure of Information
Health Canada
Public Release of Clinical Information (PRCI)
EMA
Clinical Data Publication (Policy/0070 Phase 2)
1999
2015
2018
EU
General Data Protection Regulation (GDPR)
2019
2022
TBD
EMA
Clinical Data Publication (Policy/0070 Phase 1)
EMA
EU Clinical Trials Regulation (EU CTR)

Figure 1. Timeline of key transparency regulations that came into effect for the push to publicly release clinical trial documents. Other transparency policies/guidances may apply in different regions.

Current trends emphasise the increased prevalence of public disclosure of clinical trial documents and sharing of clinical trial data (Figure 1). Additionally, forthcoming new regulations may shift and change the landscape, therefore should be monitored and incorporated into the current strategy to also achieve compliance in these areas. Transparency initiatives that publish clinical trial documents are implemented with the goal in mind of allowing public scrutiny and enabling secondary research. The scope of the GTP guideline addresses the disclosure of anonymised individual patient or participant data (IPD), both structured and unstructured, in the context of mandatory public releases as dictated by regulators and voluntary external data sharing initiatives to facilitate secondary research. The GTP guideline will use General Data Protection Regulation (GDPR) [2] terminology. However, equivalent terms from other data privacy provisions and relevant case law would apply, as the concepts are similar.

The objective of the GTP guideline is to achieve accountability and traceability through the anonymisation or de-identification process while providing reasonable assurance that privacy requirements are upheld.

1: Glossary

Many of the definitions used for the GTP have been published in a previous PHUSE deliverable – Terminology Harmonisation in Data Sharing and Disclosure Deliverables [3].

1.1 Adversary

A data user who intentionally or inadvertently learns or discloses information about a data subject through re-identification or attribution. This user may be motivated by a wish to discredit or otherwise harm the organisation disseminating the data, to gain notoriety or publicity, or to gain profitable knowledge about particular data subjects. Data adversaries are sometimes referred to as intruders, snoopers or attackers [3,4].

1.2 Anonymisation

The overall process of protecting the privacy of data subjects, including clinical study participants, and reducing the risk of re-identification by 1) modifying (e.g. suppressing, obscuring, aggregating, altering) identifiable information in individual participant data 2) assessing and controlling the residual risk of re-identification 3) considering the context of the data release [3,5].

1.3 Anonymised data and documents

Individual participant data that has been produced as the output of an anonymisation process [3,6,7].

1.4 Data Anonymiser

An entity which acts as a data processor to anonymise or de-identify individual participant data.

1.5 Data Controller

The natural or legal person, public authority, agency or other body which, alone or jointly, determines the purposes and means of the processing of personal data. Where the purposes and means of such processing are determined by union or member state law, the controller or the specific criteria for its nomination may be provided for by union or member state law [2].

1.6 Data Processor

A natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller [2]. In clinical research, the Data Processor is anyone appointed by the study sponsor to work with the clinical trial, including contract research organisations (project management, monitoring, data management, statistics, medical coding, medical writing, etc.) and vendors (eCRF/EDC, ePRO, IVRS/IWRS, central labs, etc.).

1.7 Data sharing agreement

Set out the purpose of the data sharing, cover what happens to the data at each stage, set standards and help all the parties involved in sharing to be clear about their roles and responsibilities [8].

1.8 Data subject

An identified or identifiable natural person to whom a particular piece of data relates. Depending on local legislation, a data subject may also refer to deceased individuals [3,4,9–11].

1.9 Date offset

All dates are replaced with a new date generated using an 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 [12]. An example is the 'PHUSE offset', where the offset delta for each participant is such that all participants appear to be starting the trial on the same date [5].

1.10 De-identification

A general term for any process of removing the association between a set of identifying data and a data subject present in individual participant data. The association between data and subject is removed by modifying (e.g. removing, obscuring, aggregating, altering) identifiable information in individual participant data [3,9,10,13].

1.11 De-identified data and documents

Individual participant data that has been produced as the output of a de-identification process [3].

1.12 Direct identifier

Data that can be used to uniquely identify an individual (e.g. names, initials, study participant ID, ID numbers connected with individual patient data, social security number, exact personal address, telephone number, email address, government-assigned identifier) without additional information or cross-linking other information that is in the public domain [3,9].

1.13 Generalisation

Reducing the precision of data variables. For example, aggregation can organise continuous age data into age categories, or group countries into regional or continental level [12,14].

1.14 Individual patient or participant data (IPD)

The person-specific data separately recorded for each data subject in a clinical study [3,9]. The data can be categorised as either structured or unstructured data. The distinct differences are described below to help identify which category the data fits into:

- **Structured data**
  IPD is typically presented in a dataset/spreadsheet that is easily machine readable. This dataset contains raw data exported from the study database related to a particular data point associated with an individual study participant.

- **Unstructured data**
  Often the term is used to describe the data and summary-level information found within clinical study documents. The data point can either be associated with an individual study participant or presented as aggregate-level information to summarise certain characteristics of a study population.

1.15 PHUSE Good Transparency Practice (GTP)

Guidance for the design, conduct, performance, monitoring, auditing, recording, analyses and reporting of clinical data publication or external sharing which provides assurance that the data and reported results are credible and accurate, and that the privacy rights of the trial participants are protected.

1.16 Protected personal data (PPD)

Any information relating to an identified or identifiable data subject. An identifiable subject is one who can be identified, directly or indirectly, in particular from an identification number or from one or more factors specific to their physical, physiological, genetic, mental, economic, cultural or social identity [3,15].

1.17 Primary use

Uses and disclosures that are intended for the data collected [6].

1.18 Pseudonymisation

A type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms. Typically, pseudonymisation is implemented by replacing direct identifiers (e.g. a name, a subject ID) with a randomly generated value [3,6,10,13,16].

1.19 Quantitative risk

A quantitative (numerical) approach to risk assessment uses the data from datasets or data extracted from the document (e.g. using natural language processing) or a combination of both to calculate the probability of re-identification of an individual participant and determine the amount of anonymisation that would result in the residual risk of re-identification being lower than a set threshold [17].

1.20 Qualitative risk

A qualitative (non-numerical) approach to risk assessment uses a set of transformation rules (based on rarity of disease, outliers (perceived non-standard attributes in the disclosed population), single-site trials (or other unique trial characteristics) or list of data elements that poses a risk for re-identification and uses a scale – high, medium, low – associated with each data element [17].

1.21 Quasi-identifier

Data which, in connection with other information, can identify an individual with high probability, e.g. age at baseline, race, ethnicity, gender, country, height, weight, body mass index (BMI), body surface area (BSA), genetics, medical information (also including medical history which may have occurred as a set of unique events following a unique order – which could be sensitive information with a high potential of re-identification), concomitant medications, events (also including sensitive serious adverse events), outliers in the population, family/friends information, and specific findings which are not relevant to the medical outcome [3,5,9,17–19].

1.22 Redaction

When text is obscured by an opaque box [17].

1.23 Re-identification

Re-establishing the association between a set of identifying data and the data subject found in data or documents [3,10,16,20].

1.24 Re-identification risk

The probability of re-identification occurring [3,9,10,21].

1.25 Residual risk

The risk of re-identification of an individual participant, which remains based on the data variables disclosed in datasets or documents that have been produced as the output of an anonymisation process [3,9,21].

1.26 Risk threshold

The maximum amount of acceptable re-identification risk remaining in documents and data after an anonymisation process has been applied. The threshold value can be either quantitative or qualitative [3,22].

1.27 Secondary use

Uses and disclosures that are different from the purpose(s) for which the data was collected as described in a clinical trial protocol and informed consent form [3,6].

1.28 Sensitive information

Any data which, in the event of re-identification, could be considered harmful for a data subject in terms of employability, reputation, insurability, self-esteem or stigma, or could result in loss of income. The perception of information as sensitive is subjective, and examples include genetic information, substance abuse, mental disorders and abortion [3,9].

1.29 Single out

To isolate some or all records which identify a data subject in the dataset by observing a set of characteristics known to uniquely describe that data subject [3,23,24].

1.30 Suppression

When the text is removed (and potentially replaced with other text and/or special signs). Suppression may still allow some information to be considered and assessed [17].

2: The Principles of PHUSE GTP

2.1 Clinical trial data sharing should be conducted in accordance with the applicable regulatory requirement(s) and local legislature [1,21,25].

2.2 Anonymisation should adequately protect participant privacy and prevent a serious possibility of them being identified in the data and/or associated with information that should be protected to keep them anonymous [21,25].

2.3 Prioritise retaining the maximum amount of analytically valuable information during the anonymisation process while maintaining patient privacy [21,25].

2.4 The chosen anonymisation strategy should be documented in an anonymisation report made available to the Data Recipient.

3: Data Controller

A Data Controller is the entity that will determine the purposes of collecting personal data and the means of processing said personal data [2]. The specific criteria determining which party will assume the responsibility of Data Controller is determined by local law and, in the case of outsourcing of data processing, is further refined by data protection agreements. In the context of clinical trials, this role will usually belong to the trial sponsor. Other organisations may qualify as joint controllers, e.g. a contract research organisation (CRO) that has been delegated a full clinical development responsibility, or a primary investigator in an academic trial.

3.1 Responsibilities

The Data Controller has multiple key responsibilities to protect participant data collected during a clinical trial, including [2]:

- Liable to data subjects for non-compliance of data protection regulations (depending on regional guidelines)
- Ensuring the lawful collection and processing of the data
- Establishing and maintaining responsible sharing for reuse purposes to further science
- Providing safe storage and disposal of personal data
- Reporting data breaches
- Controlling access to personal data and emphasising the best practice for data recipients to adhere to the terms of use.

3.1.1 Transparency quality assurance

Beyond the responsibilities outlined by the GDPR, under PHUSE GTP the Data Controller should also be responsible for ensuring the:

1. Privacy of human participants is protected.
2. Anonymised data is accurate, legible, and has been processed correctly.
3. Anonymised data has been processed using best practices outlined by the PHUSE GTP and all other applicable regulatory requirements.

There are key documentation procedures and governance frameworks which could be implemented to ensure these responsibilities are met, such as:

- Standard operating procedure (SOP) and work instruction (WI) creation and implementation
- Recordkeeping
- GTP internal monitoring committees
- GTP stewardship councils
- Audits.

PHUSE GTP is an external guidance which data controllers can use in conjunction with other guidances. Each organisation has its own governance and bodies responsible for complying with company procedures and business process owners.
The model outlined below may be used by data controllers to create a framework to govern data sharing. Depending on the number of activities, similar processes may be carried out by existing teams, such as a clinical trial transparency or a data management group.

SOP and WI creation and implementation

The Data Controller (potentially in collaboration with the Data Anonymiser, if they are a separate party) is responsible for writing SOPs and WIs which define, for example, the:

1. Process for anonymisation of clinical trial documents (including roles and responsibilities of Data Anonymiser personnel, the method chosen, and the tools used to perform the anonymisation)

2. Process for anonymisation of clinical trial datasets (as above)
3. Quality control procedures in place to ensure the appropriate identifiability thresholds are not exceeded and the anonymised data or documents remain fit for purpose (including any certification/assurances)
4. Storage and transmission of clinical data and documents before and after the anonymisation has been applied, including retention periods (especially where the Data Anonymiser is a separate party).

SOPs should be maintained and updated regularly.

Recordkeeping
Similarly, the Data Controller must create and maintain:

1. Records of data (structured or unstructured) which have been anonymised, including dates of sharing/publication, the governing transparency requirements, data recipients and data retention periods external to the Data Controller organisation
2. Relevant personnel within the Data Controller teams who have been trained to the degree commensurate with their role in:
   a. Legal requirements and international best practices in clinical trial document disclosure, clinical data transparency and data sharing
   b. The contents of the SOPs (outlined above).

Training refreshers may be required periodically.

Records must be maintained, signed by the relevant personnel (e.g. compliance officers) and kept up to date. These records may be called into review during periodic audits (detailed below).

GTP Internal Monitoring Committees
The purpose of the GTP Internal Monitoring Committee is to ensure:

1. Anonymisation and subsequent sharing/publication of clinical trial data assets is being completed to the standard outlined in the SOPs, and to maintain the privacy of the represented clinical trial participants
2. The SOPs remain applicable and reflect the current regulatory environment and anonymisation best practices
3. All records have been kept up to date
4. Any other items relevant to adherence to GTP are discussed as appropriate.

The Internal Monitoring Committee should contain representatives from the Data Controller organisation but may require some input from the Data Anonymiser on item 2.

This committee should prepare an Internal Monitoring Committee report for sharing with the GTP Stewardship Council.

GTP Stewardship Council
The role of this council is to oversee the Internal Monitoring Committee and to align on any strategic or organisational changes made in response to the findings of the Internal Monitoring Committee.

Audits
The purpose of the audit is to ensure that data and documents are anonymised in keeping with the SOPs, that the tools used to perform the work remain fit for purpose, and that the appropriate standards and regulations are being adhered to by both the Data Controller and the Data Anonymiser. The auditor should be selected by the Data Controller and be appropriately qualified and independent of the clinical trial transparency function.

The auditor may seek input from the Data Anonymiser if separate from the Data Controller.

The observations and findings of the auditor(s) should be documented.

3.2 Role

The Data Controller may perform anonymisation on the data themselves or outsource to a Data Processor – a third-party organisation – to process the data. In the context of the PHUSE GTP, we will use the term Data Anonymiser to refer to Data Processors who perform anonymisation services. The Data Controllers are responsible for creating data processing instructions for the Data Anonymiser. A data processing agreement (DPA) will detail how the Data Anonymiser will process and store data, including how long for. The planned anonymisation strategy and process flow will also be described in this contract. The DPA may also delegate some of the Data Controller's responsibilities, and related liability, for patient re-identification on to the Data Anonymiser.

3.2.1 Anonymisation for regulatory compliance

Data Controllers may be required to make anonymised clinical documents publicly available to comply with multiple health authority transparency policies/guidances (Table 1). In the future, anonymised clinical trial structured IPD may also be required to be publicly available under policies such as Phase 2 of the EMA Clinical Data Publication (CDP), also known as Policy 0070 [21]. Clinical trial documents may also be requested by individuals under policies such as the Freedom of Information Act [26] or the EMA Access to Documents, also known as Policy 0043 [27], or other regional equivalents. Please note this is not an exhaustive list and it captures the most prevalent, well-known initiatives in the industry.

Regulators prefer submissions for the EMA CDP [21] and Health Canada Public Release of Clinical Information (PRCI) [25] to use a quantitative risk-based anonymisation strategy rather than a qualitative redaction-based strategy; however, both may be accepted. Anonymisation strategies should demonstrate how their chosen methodology results in a re-identification risk equal or lower than the default threshold of 0.09 [21,25,28]. Public releases, such as regulatory submissions, should as best practice use a 'maximum risk' approach, as there are minimal controls of who may access the data and no time limit on the access to the data [20]. A maximum risk approach measures the risk of re-identification as the maximum value across all participants [20].

Table 1. Summary of key policies/guidances that publicly release clinical trial documents. Other transparency policies/guidances may apply in different regions.

Region	Health Authority	Title	Description
Japan	Pharmaceuticals and Medical Devices Agency (PMDA)	Disclosure of Information [29]	Clinical trial documents (not including the full clinical study report) will be publicly posted on the PMDA website [30].
European Union	European Medicines Agency (EMA)	Clinical Data Publication (CDP), also known as EMA Policy 0070 [21]	Clinical trial reports (Phase 1) submitted under the centralised marketing authorisation procedure will be publicly posted. Clinical Data Publication portal [31] launched for Phase 1 in 2015; Phase 2 scope and implementation date is currently unknown.
European Union	European Medicines Agency (EMA)	EU Clinical Trials Regulation (EU CTR) No. 536/2014 [32]	Clinical trial information and documents during the life cycle of a clinical trial will be publicly posted. This includes investigational medicinal products regardless of whether they have a marketing authorisation. Clinical Trial Information System (CTIS) [33] launched in 2022.
Canada	Health Canada	Public Release of Clinical Information (PRCI) [25]	Clinical trial documents in applications that have been authorised, including those authorised and then revoked, will be publicly posted. Clinical Information portal [34] launched in 2019.

3.2.2 Anonymisation for voluntary data sharing

Sponsors may also participate in voluntary data sharing to facilitate secondary research on platforms such as Vivli [35], ClinicalStudyDataRequest.com [36] or the Yale University Open Data Access (YODA) Project [37]. See a comprehensive list of data sharing platforms compiled by Yale University [38].

The appropriate anonymisation strategy for voluntary data sharing needs to be determined on a case-by-case basis, depending on the context. The re-identification risk threshold value is at the discretion of the Data Controller [20]. Voluntary data sharing typically has increased data controls than the public release of data. These increased data controls can include limiting access to legitimate researchers who have signed a data sharing agreement, restricting viewing access to a secure portal with no local download option and/or timebound access to the data. These increased data controls may be used to justify using a less conservative anonymisation strategy which can increase data utility. The Data Controller may consider choosing a re-identification threshold higher than 0.09 and smaller equivalence classes. Another less conservative strategy, such as an 'average risk' or a 'strict average risk' approach, can be sufficient to protect patient privacy in private releases [20]. An average risk approach is where the risk of re-identification is the average value across all patients, whereas a strict average risk approach is the average value across all patients and requires no value to be above a predetermined threshold.

In voluntary data sharing, data utility is critical, as secondary research proposals may have certain variables of interest. If resources allow, the Data Controller should survey which variables are necessary for the researcher's proposal and communicate this to the Data Anonymiser so that the anonymisation strategy can prioritise retaining these variables. If multiple variables are important, the researcher should be asked which are their top priority.

3.2.3 Optimising data for anonymisation

In clinical trials, the Data Controller is responsible for the creation of all trial records. These include:

- study records
- case report forms (CRFs)
- electronic database captures (EDCs)
- safety databases
- electronic trial master files (eTMFs)
- data at the end of the studies for publication
- analysable datasets for sharing with other parties
- clinical study reports (CSRs) for submission to regulatory authorities
- other documents as required (e.g. protocol summary, results summary, plain language summary).

The need for potential anonymisation for all future clinical trials should be anticipated by the Data Controller when formatting the original data, to make processing as efficient as possible. Structured IPD should abide by the standards described by the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) [39] and the Analysis Data Model (ADaM) [40] or any other standards that apply and have sufficient supporting documentation. For studies which took place before these standards were created, there should be an effort to provide documentation of the data and its structure. Clinical trial results should be presented according to the format described in International Council for Harmonisation (ICH) E3 guidelines [41] for formatting clinical study reports and strive to be clear, consistent and concise. For specific examples of best practice suggestions, see Appendix 1.

4: Data Anonymiser

Data anonymisation is a processing activity that considers the context of protected personal data (PPD), assesses and controls residual risk of identifiability, and removes or modifies information to anonymise data that cannot be associated with any one individual.

The Data Controller may be the Data Anonymiser, or the Data Controller may choose to engage a third party to process the data. If so, a DPA outlines the roles and responsibilities of the new data processor, based on which they may be liable to data processing regulations.

4.1 Responsibilities

A Data Anonymiser is responsible for identifying and transforming (or removing) personal identifiers from data while ensuring the data remains useable for its intended purpose. Their main responsibilities include:

- Understanding and interpreting regulations, governance laws and industry standards related to data privacy and anonymisation
- Implementing and maintaining an appropriate anonymisation strategy, as instructed by the data controller, which fulfils both the Data Controller's objectives and regulatory requirements
- Confirming the anonymised data has been processed in compliance with regulatory guidelines and internal SOPs
- Ensuring the safe storage of data by:
    - Controlling access to data
    - Processing personal data only as intended by the Data Controller and keeping records thereof
    - Reporting data breaches to the Data Controller and any additional governance committee, and cooperating with local data protection authorities
    - Deleting/returning personal data after the service contract is complete per the DPA, if applicable
    - Employing a data protection officer and/or EU representative, if appropriate.

4.2 Role

4.2.1 Implementing an anonymisation strategy

The Data Anonymiser should have a robust understanding of the methodology for anonymisation practices and should process data in compliance with the de-identification strategies provided by the Data Controller while ensuring compliance with local regulations and guidelines. In practice, the anonymisation strategy is devised by the Data Controller, with the data processor being responsible for implementing it. The implementation of a suitable anonymisation strategy as endorsed by the Data Controller includes:

1. Identifying and classifying direct and quasi-identifiers from datasets and determining whether personal information is linkable throughout the source data or through additional data that may be available to the Data Recipient.

2. Determining an anonymisation strategy that is compliant with GDPR [2], the Data Controller's organisational policies, and local regulations. The Data Anonymiser should consider the prioritised data variables highlighted by the Data Controller to maximise data utility. The following factors to consider when determining an appropriate anonymisation strategy include:

    a. Measuring the identifiability of personal information and applying appropriate mitigation strategies to anonymise the data properly. Identifiability risk can be quantified by the following equation [42]:

    P(identification) = P(identification|threat) x P(threat)

    Note: The conditional probability P(A|B) is the probability of A given B, or the probability that A will occur on the condition that B occurs.

    b. Evaluating additional data available, which might lead to inappropriate revealing or unmasking data that should be protected.

    c. Assessing the risk of an attack by an adversary which would reveal personal information (deliberate, accidental, environmental, etc.).

    d. Determining the reference population, either from participants in the clinical trial being studied or through an extrapolation method from other data sources.

    e. Considering the purpose of the data, its context and its intended sharing environment (i.e. structured data vs unstructured data, internal sharing vs external sharing, regulatory vs voluntary sharing), while considering additional data sources available. In public sharing, since the data recipients are members of the general population and are not subject to any DSA, this poses the highest risks.

    f. Consistency of similar data protection used across multiple releases where possible, to not exceed the recommended threshold by linkage, which could lead to methods disruption or inadvertently revealing original values determined through comparison analysis or from context.

    g. Providing justification on how the chosen anonymisation method achieves regulatory requirements (e.g. the Safe Harbor Method vs the Expert Determination Method).

3. Applying appropriate transformations to the data and producing an anonymisation report. The anonymisation report details how the risk of re-identification has been measured, describes the applied anonymisation technique and outlines how the data is to be shared. The Data Controller reviews and endorses the anonymisation report to ensure the report accurately represents the implemented anonymisation technique and follows the organisation's data protection policies. The anonymisation report allows external parties, such as regulators, to assess the effectiveness of the anonymisation strategy and ensure compliance with relevant data protection laws and regulations [5,21].

4.2.2 Applying anonymisation techniques

When applying a particular anonymisation strategy, the Data Anonymiser must consider local regulations and policies, the purpose of the data, and its intended sharing environment.

Regulatory agencies such as the EMA CDP and Health Canada PRCI have stated a preference in the use of quantitative risk assessment over qualitative, a goal that aligns closely with GDPR [21,25]. The implementation of a quantitative risk assessment is a statistical approach which involves determining an appropriate risk threshold for re-identification, calculating the probability of re-identification and applying a strategy that allows this risk to fall beneath a set risk threshold. Both the EMA and Health Canada recommend the risk threshold not exceed 0.09 for documents that are made publicly available (Figure 2).

Conversely, regulatory agencies accept submissions that use a qualitative risk assessment, which is a rules-based approach based on the identifiability of characteristics from the data source (i.e. rarity of disease, number of patients and sites in the study) and uses a qualitative scale (high, moderate, low) to classify risk [9]. *Note: Qualitative risk assessments are not supported by the literature, and there is no evidence of their efficiency.* Examples of how data may look after transformation can be found in Appendices 2 and 3.

Additional anonymisation techniques include:

- Pseudonymisation: when the Data Anonymiser replaces direct identifiers with encrypted data or false identifiers (refer to Table 2 for example).
- Generalisation: when the Data Anonymiser replaces a data value with a generalised term or group (e.g. age/BMI/weight/height can be banded into ranges, and locations can be generalised to global regions) (refer to Table 2 for example).
- Suppression: removing or eliminating information by replacing text with an unrelated value or text.
- Redaction: removing data entirely by obscuring text with an opaque box. If the regulator is associated with another entity whereby there may not be a specific guideline to follow, the default method for the redaction overlay text would be to follow the approach described by the EMA.
- Date offsetting: offsetting key participant-specific date variables across the study (i.e. randomisation date, informed consent date, study start date, medical history date). Dates can be offset to an anchor date via PHUSE date-shifting, where the anonymised data reflects that every participant started the trial on the same day, and individual intervals between dates are maintained [18]. Conversely, random offsetting is where each participant is given a random offset that is applied consistently to all dates for that participant. Other date variations, such as partial or imputed dates, must be carefully considered, as they may reveal additional information and impact on the risk assessment [46].

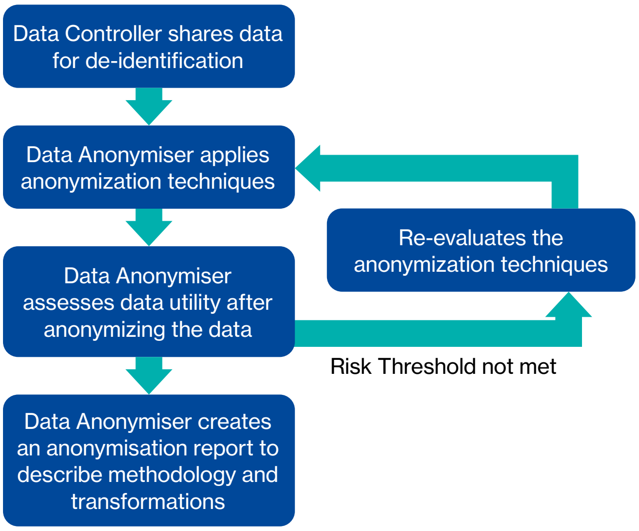

Data Controller shares data for de-identification
Data Anonymiser applies anonymization techniques
Data Anonymiser assesses data utility after anonymizing the data
Re-evaluates the anonymization techniques
Risk Threshold not met
Data Anonymiser creates an anonymisation report to describe methodology and transformations

Figure 2. Application of an outsourced anonymisation strategy following the assessment of re-identification risk. The anonymisation strategy is selected by the Data Controller and applied by the Data Anonymiser. For an in-house anonymisation strategy, the Data Controller and the Data Processor are different depending on the internal company team structure.

Subject ID	Age	Date of Birth	Sex	Country	Race	Informed Consent Date
PROTOCOL01-US001-001001	26	1995-04-01	M	USA	WHITE	2019-04-19
PROTOCOL01-US001-001002	36	1985-03-01	F	USA	WHITE	2019-04-02
PROTOCOL01-US002-001003	20	2001-04-05	M	USA	MULTIPLE	2019-03-22
PROTOCOL01-US003-001004	22	1999-04-05	F	USA	BLACK OR AFRICAN AMERICAN	2019-03-19
PROTOCOL01-CA001-001001	31	1990-01-23	M	CAN	WHITE	2019-03-16
DeID Subject ID	DeID Age	DeID Date of Birth	Sex	DeID Country	DeID Race	Informed Consent Date
PROTOCOL01-FP195-843195	(25–29)	NULL	M	North America	WHITE	2019-03-15
PROTOCOL01-ZR310-066432	(35–39)	NULL	F	North America	WHITE	2019-03-15
PROTOCOL01-GP023-332941	(20–24)	NULL	M	North America	NULL	2019-03-15
PROTOCOL01-MK329-278395	(20–24)	NULL	F	North America	BLACK OR AFRICAN AMERICAN	2019-03-15
PROTOCOL01-PE082-382816	(30–34)	NULL	M	North America	WHITE	2019-03-15

Table 2. Example of source data (top) and anonymised data (bottom) which demonstrates Subject ID pseudonymisation, generalisation by providing age banding and generalised countries to global regions, date of birth and race suppression, original sex and race values retained, and date-shifting. Note: The term 'De-Identification' is shortened to 'DeID'. See Appendix 3 for additional dataset examples of anonymisation strategies. Please note the examples are not an exhaustive list and highlight the most frequently observed scenarios.

4.2.3 Special conditions to consider for the anonymisation strategy

Following the application of a chosen anonymisation strategy, the Data Anonymiser should consider the possibility of residual risk, which is the risk attributed to re-identification of anonymised data through additional analytical techniques or by combining the data with supplementary information. Aggregated data should be carefully analysed when determining the anonymisation strategy, as summary-level information such as subgroup analysis by age, sex and region may impact on risk [43]. Through further manipulation of the data, an attacker may single out an individual participant by isolating data records that could characterise them. This potential risk of re-identification can be mitigated by carefully analysing the dataset to determine if there are any unique characteristics or events associated with the participant. Such characteristics should be incorporated into the anonymisation strategy for the protection of these highly identifiable attributes. For instance, a serious adverse event that might be sensitive and requiring redaction (e.g. suicide attempt, amputation) should be informed by the risk assessment if considered a quasi-identifier, or by the risk threshold if considered sensitive.

It is also possible for the summary-level information to unmask redactions on IPD. For example, assuming 2 arms A and B (from a summary demographic table, we know that arm A = 8 participants – 8 white, arm B = 10 participants – 8 white, 2 Asian). For this example, the risk simulation could be Race = Drop (due to 2 Asians). Due to this relatively small number, the approach may be to suppress Race in the IPD. However, from the summary level, we know that all patients (8) in arm A are white, thus if we find a patient from arm A with redacted race, we know from the summary table that the patient is white. Therefore, to mitigate this risk, both instances (individual and summary level) could be considered for protection in the anonymisation strategy.

The Data Anonymiser should therefore consider such summary-level information from the clinical study report in addition to results posted in public registries, which may build a comprehensive patient demography profile. Therefore, to mitigate potential linkage and to protect privacy prior to public release, it is essential to cross-reference all information available and target those aspects to secure the information and eliminate this risk. Please note not all risk measurements will have this outcome. For example, the reference population might be very large if this was a healthy participant study and race can be retained.

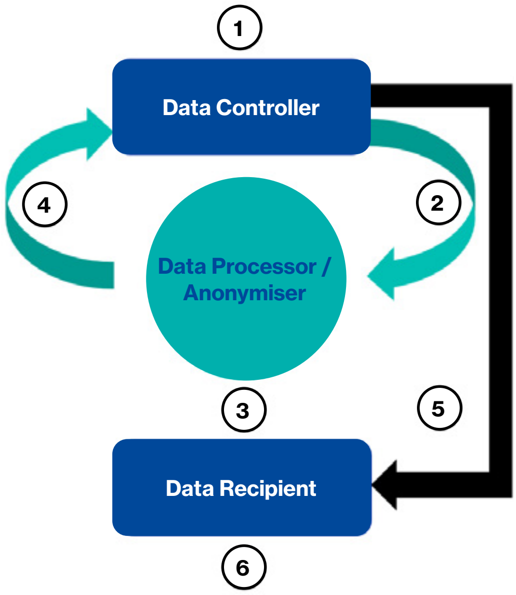
Although a risk threshold may be met for a particular dataset, atypical values or special characteristics within the dataset could still be present and pose a risk of identification. Threshold reflects the sensitivity and potential invasion of privacy, not special cases in the data. Conditions such as rare disease studies, low-frequency events and demographic characteristics, or planned analyses for population sub-groups, could provide the public with additional information for identifying trial participants and should be considered when implementing an anonymisation strategy. Country or region-specific drug marketing names should also be considered to avoid compromising the generalisation of geographical location. Additionally, consideration should be given when processing datasets that include sensitive information, genetic data, seasonal data, or images [46]. Such conditions should be treated on a case-by-case basis to ensure personal information is adequately anonymised.

Finally, the Data Anonymiser should be mindful of risks and threats posed by the continuous development of artificial intelligence (AI) within the healthcare industry. As AI capabilities continue to expand within today's clinical landscape, it is important to recognise that de-identification methodologies and privacy regulations risk falling behind the principles they govern. Although strategies such as anonymisation help safeguard patient data, future AI threats may create vulnerabilities that the Data Anonymiser may not yet consider. As the field of AI continues to grow, the Data Anonymiser may also want to assess what other data the user has access to as part of the controls and settings that need to be evaluated, particularly with open-access uncontrolled regulatory sharing. Adversarial modelling and risk assessments should be updated regularly to reflect new AI techniques and data sources in an effort to mitigate new AI-driven re-identification threats [47].

5: Data Recipient

The Data Recipient is the individual or group receiving anonymised data from the Data Controller. This data will have been anonymised by the Data Processor.

The Data Anonymiser will anonymise the required data before sharing it with the Data Recipient to protect the privacy of trial participants, study personnel and company confidential information associated with the study sponsor.



Key:

[1] The Data Controller is responsible for the data collected at the clinical trial study site.

[2] The Data Controller instructs the Data Processor to anonymise the data.

[3] The Data Processor uses the appropriate anonymisation technique depending on the purpose of sharing the data.

[4] The anonymised data is shared with the Data Controller for review and approval.

[5] Approved anonymised data is sent to the Data Recipient from the Data Controller.

[6] The Data Recipient upon receiving the anonymised data uses the data for their purposes and will abide by the responsibility to not attempt re-identification.

Figure 3. Relationship between Data Recipient, Data Controller and Data Processor in an outsourced model of anonymisation (for example between the study sponsor pharmaceutical company who would be the Data Controller and the Data Processor; a CRO/anonymisation vendor who anonymises the data). Please note, in the case of shared data sharing platforms, access to the data is not generally controlled by the Data Controller, but by the administrators of the data sharing platform. Another note to consider is for internal models, whereby the Data Processor/Anonymiser are different depending on the internal company team structure.

Data from clinical studies can be shared in one of two forms, either structured data or unstructured data. Currently only unstructured data (e.g. documents) are made publicly available under transparency initiatives. However, in a private sharing scenario, Data Recipients may be given access to one and/or both.

5.1 Responsibilities

The Data Recipient's responsibility when handling anonymised data depends on whether they participate in private or public data sharing.

Access to data by recipients can be grouped into two main categories: Private Sharing and Public Sharing. Both categories relate to who will ultimately be the end user (i.e. who will have access to the data).

5.1.1 Private Sharing

In Private Sharing, the recipients will sign a data sharing agreement (DSA) and are contractually obligated to uphold the confidentiality obligations of the data participants.

As described below, the DSA is a formal contract which is not always between the Data Controller and the Data Recipient, typically a qualified external researcher. It provides an in-depth overview of how the data can be used, and the legal rights and obligations of the Data Controller and the Data Recipient when a sponsor company (Data Controller) shares the anonymised data. The process for Private Sharing generally involves liaison with the other parties, in which case the Data Controller does not always grant access to the data. For Vivli and YODA for example, the DSA is between Vivli/Yale University and the Data Recipient. The DSA will always include a clause requiring the Data Recipient to agree not to attempt the re-identification of participants in the anonymised dataset. The DSA would also enforce the Data Recipient to inform the Data Controller regarding any breach on their end.

Note: A draft template DSA is typically created by the other party (i.e. CSDR.com, YODA or Vivli) and shared with the Data Recipient, whereby both parties review and negotiate the terms to reach a mutual agreement for the legal wording. In instances where the request is received outside this channel, the Data Controller may already have their own company draft template DSA in place to be shared with the Data Recipient. The DSA is considered a fully executed binding contract once all parties involved have signed the document. Only once the DSA is fully executed will it be permissible for the Data Controller to grant data access to the Data Recipient. Please note, as described in the above example for Vivli and YODA, the Data Controller does not always grant access to the data.

5.1.2 Public Sharing

Generally, in Public Sharing, the Data Recipients are members of the general population and are not subject to any DSA. Although not as enforceable as a DSA, the terms of use here and in any form of public disclosure are considered to be legally binding. Data Recipients will be able to access anonymised clinical documents for non-commercial purposes. An example would be a member of the public visiting the European Medicines Agency (EMA) or Health Canada (HC) Clinical Data Publication portal. Please note only citizens of the European Union can download from the EMA's CDP portal. This uncontrolled data sharing poses the highest risks, therefore Data Recipients accept the terms of use in this clinical data publication context (when data is released into the public domain vs private/individual sharing with a trusted partner or organisation).

Under the terms and agreements of use, the Data Recipient agrees to not attempt re-identification of participants in the anonymised dataset.

5.2 Role

The main three types of Data Recipients are summarised below, along with the purposes for which they require anonymised data:

- **Researchers**

Generally, for this guidance, researchers include qualified external requestors engaged in independent scientific research who require access to structured and/or unstructured data to perform research on their topic of interest under a research proposal to test their hypothesis(es) or answer their research question(s).

- **Regulators**

An independent entity with the purpose of developing, enforcing and monitoring guidelines to ensure a product or service meets the highest standards.

- **Public**

Any member of the general population, regardless of their sector or profession.

To understand the anonymised data received, Data Recipients must have the necessary background knowledge to help them interpret and use the information. Incorrect interpretation gives rise to misunderstanding and the spread of misinformation, therefore a good understanding by Data Recipients is essential.

The level of detail in understanding the data depends on how a Data Recipient wishes to use the data.

Summarised below are the recommended areas for Data Recipients to understand before using and analysing the anonymised data, along with where to obtain the tools, resources and training available in these areas.

5.2.1 General understanding of data privacy and data sharing in clinical trials

There is increasing interest in data privacy and data sharing, and most publicly available information is targeted at individuals with training, experience or expertise in these areas. It is important to have a good understanding of data privacy and data sharing in clinical trials to create awareness around how data is used and the rights, policies and benefits for individuals. This also helps tackle the spread of misinformation and misconceptions in the public domain. The PHUSE Project Educating the General Population on Data Privacy and Data Sharing [44] focuses on creating engaging content organised into short videos on data privacy and data sharing which can be understood and used by the general population (any member of the public regardless of their sector or profession). These videos cover commonly asked questions and will increase general public knowledge (and accuracy thereof), by providing accessible and understandable content. Ultimately, this understanding will enable the viewer to become more informed, thus having a positive impact on their families and the wider community.

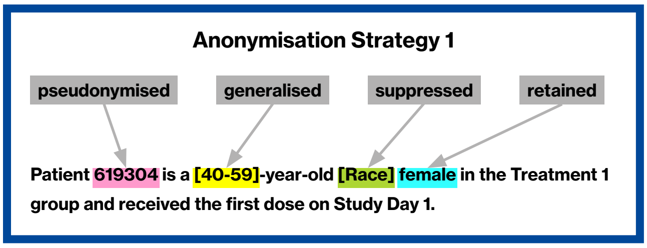
5.2.2 Knowledge of anonymisation process and techniques

The Data Recipient should be familiar with different anonymisation strategies and techniques (see examples below). Other examples of unstructured data anonymisation are found in Appendices 2 and 3, and examples of structured data anonymisation are in Appendix 4. They should consider how anonymisation methods and, therefore, data utility may change in contexts (structured data vs unstructured data, regulatory vs voluntary). The 'context' is about how the data is shared and what measures are in place to control sharing. Finally, the Data Recipient should consider how the anonymisation strategy has transformed the data to analyse the data correctly and not to mislead the results. Incorrect anonymisation would result in incorrect results. For example, if dates are offset, an analysis of seasonality would be misleading since the anonymised dates might not be in the same season as the original data.

Anonymisation examples

Original narrative text:
"Patient 001002 is a 49-year-old white female in the Treatment 1 group and received the first dose on Study Day 1."

More than one anonymisation strategy can achieve a re-identification risk lower than the acceptable threshold.
The following text shows how a robust numerical approach is used to determine the risk of re-identification based on variables such as population size and demographics. This approach will form the rationale for how the personal identifiers are handled. This example showcases a wealth of different aspects of what an anonymised document might look like, using four potential ways of anonymising the sentence:


Anonymisation Strategy 1

Patient 619304 is a [40-59]-year-old [Race] female in the Treatment 1 group and received the first dose on Study Day 1.

Please refer to Appendices 2 and 3 to see additional anonymisation examples.

There are current efforts by regulators and study sponsors to share clinical study reports (CSRs) and IPD from clinical trials more widely. The PHUSE De-Identification Working Group project focused on defining de-identification standards for CDISC standards and was released in 2015 as the PHUSE De-Identification standard for SDTM 3.2 [18]. The goal is to define standards to reduce efforts for companies to anonymise IPD and provide consistent data to researchers where data utility is considered.

Other factors to consider on a case-by-case basis when determining risk of re-identification are the population size, demographics (such as only one individual in a race or gender category) and whether the condition being studied is a rare disease. It is also important to note that there can be a wider spectrum of this anonymisation approach to capture additional elements, such as when medical history and type of event is accompanied by System Organ Class for which the narrative was created. The same data may be presented across multiple sections across the clinical study report (such as tables, listings, and baseline summary characteristics), which should be taken into consideration when determining the anonymisation strategy to be applied in the narratives.

The overall focus of this guidance is on anonymised data. Confidential commercial information (CCI) or confidential business information (CBI) is obscured entirely with an overlay text and redaction box depending on which regulation the clinical data publication is being released (for example, EMA [21,31] or HC [25,34]).

5.2.3 Familiarity with anonymisation reports

Please refer to the joint EMA and HC template available online from the EMA website via: Home > Human regulatory: overview > Marketing authorisation > Clinical data publication > Support for industry on clinical data publication. It is important to keep in mind these templates are designed for the public sharing of anonymised unstructured data and/or documents and would not be appropriate to support anonymised structured data. The uses of an anonymisation report will vary depending on the Data Recipient type:

- **Researchers**

  Refer to the anonymisation report to understand the anonymisation methods used and have a particular interest in the variables related to their research topic. The researchers do not typically submit a research proposal to the study sponsor (data owner) but to the external data sharing body, as independence is important when expressing their interest in conducting secondary analyses for research purposes.

- **Regulators**

  Review the anonymisation report to understand the anonymisation methodology and how this will impact on data utility. The regulators can use this review to provide recommendations to the study sponsor (data owner) on how to maximise data utility and the most effective strategy to share data effectively whilst upholding the highest privacy standards in compliance with privacy laws.

- **Public**

  Reading through the anonymisation report will be useful. Although the general population may not be trained in the field of data anonymisation, the document provides background information which will help them understand the methodology used.

5.2.4 Methods of receiving anonymised data

The diagram/infographic below summarises increasing levels of protection in how data and documents are received based on the level of identifiability, context and risk of re-identification [45].

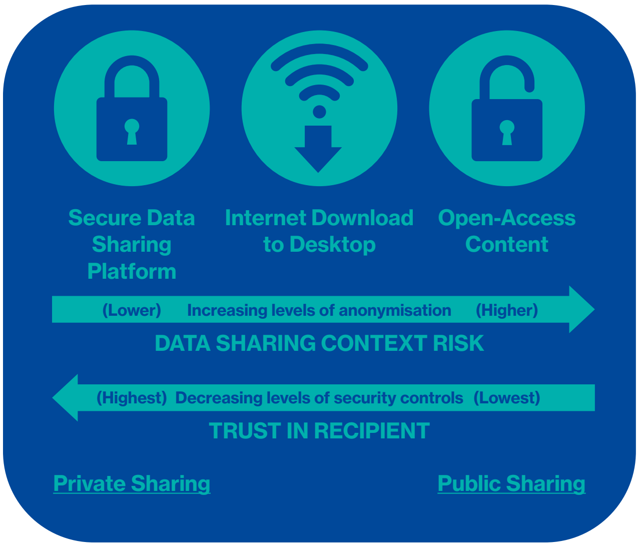

Secure Data Sharing Platform
Internet Download to Desktop
Open-Access Content
(Lower) Increasing levels of anonymisation (Higher)
DATA SHARING CONTEXT RISK
(Highest) Decreasing levels of security controls (Lowest)
TRUST IN RECIPIENT
Private Sharing
Public Sharing

Figure 4. Comparison in the varying levels of data sharing methods and anonymisation required, depending on the trust in recipients in Public Sharing vs Private Sharing.

Data identifiability can vary depending on the setting in which it is shared. When data is accessed, and the analyses are performed within a secure platform such as CSDR, YODA or Vivli, this setting can be referred to as a Secure Data Sharing Platform in which the context of risk and stringency of anonymisation is at a lower level. A medium level of context risk and stringency of anonymisation would apply to data that can be downloaded to a local machine from a data sharing platform, such as Vivli's data download functionality, which will be referred to as the Internet Download to Desktop option.

The difference between the Secure Data Sharing Platform and Internet Download to Desktop option is that the latter is no longer 'secure' since the data is outside the platform and could be freely distributed. This lack of security could cause issues if the Data Recipient failed to adhere to their contractual agreement (e.g. DSA) to uphold the privacy and anonymity of the participants. Finally, Open-Access Content is the uncontrolled data environment that poses the highest context risk since the data is freely available to the public, such as those under the EMA CDP and HC PRCI online portals, and therefore requires the highest levels of stringency for anonymisation. The Secure Data Sharing Platform setting is the most common method used in the context of Private Sharing, and Open-Access Content is typically seen in Public Sharing.

When comparing the settings from the Secure Data Sharing Platform to the Internet Download to Desktop option and finally to Open-Access Content, the security controls decrease, and so does the trust of the data being misused, therefore the stringency of anonymisation increases. Patient privacy is of paramount importance and so the greatest level of due diligence is required for data and documents to be used for Public Sharing. Balancing the extent to which data is anonymised with the context risk of a given sharing scenario will maintain the maximal level of patient privacy and data utility.

6: References

1. International Council for Harmonisation: ICH E6 (R3) Guideline for Good Clinical Practice (GCP). (06 January 2025). Accessed at: https://database.ich.org/sites/default/files/ICH_E6%28R3%29_Step4_FinalGuideline_2025_0106.pdf

2. European Medicines Agency: EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679. (27 April 2016). Accessed at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

3. PHUSE: Terminology Harmonisation in Data Sharing and Disclosure Deliverables: Terms and Definitions, Version 2.0. (2 November 2021). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/WP065.pdf

4. Elliot, M., Mackey, E., O'Hara, K. et al. The Anonymisation Decision-Making Framework (2016). UK Anonymisation Network. Accessed at: https://eprints.soton.ac.uk/399692/1/The-Anonymisation-Decision-making-Framework.pdf

5. PHUSE: Data Anonymisation and Risk Assessment Automation, Version 1.0. (9 June 2020). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+Anonymisation+and+Risk+Assessment+Automation.pdf

6. International Organization for Standardization: ISO 25237:2017(en) Health informatics – Pseudonymization. (January 2017). Accessed at: https://www.iso.org/obp/ui/#iso:std:iso:25237:ed-1:v1:en

7. International Organization for Standardization: ISO/IEC 29100:2024(en) Information technology – Security techniques – Privacy framework. (February 2024). Accessed at: https://www.iso.org/standard/85938.html

8. Information Commissioner's Office. Accessed at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/data-sharing-a-code-of-practice/data-sharing-agreements/

9. PHUSE: Protection of Personal Data in Clinical Documents – A Model Approach, Version 1.0. (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Protection+of+Personal+Data+in+Clinical+Documents+A+Model+Approach.pdf

10. Garfinkel, S. L. (October 2015). 'De-Identification of Personal Information'. Internal Report 8053. National Institute of Standards and Technology. Accessed at: http://dx.doi.org/10.6028/NIST.IR.8053

11. International Association of Privacy Professionals: Glossary of Privacy Terms. Accessed at: https://iapp.org/resources/glossary (last accessed 22 March 2021).

12. PHUSE: De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach, Version 1.0. (10 June 2019). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+and+Anonymization+of+Individual+Patient+Data+in+Clinical+Studies+a+Model+Approach.pdf

13. Clinical Data Interchange Standards Consortium: Glossary, V15.0. (18 December 2020). Accessed at: https://www.cdisc.org/standards/glossary

14. El Emam, K., Rodgers, S. & Malin, B. (2015). Anonymising and sharing individual patient data. BMJ, 350:h1139.

15. European Union: Directive 95/46/EC (Data Protection Directive). (24 October 1995). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046

16. National Institute of Standards and Technology: Computer Security Resource Center Glossary. Accessed at: https://csrc.nist.gov/Glossary (last accessed 22 March 2021). National Institute of Standards and Technology: NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0. (16 January 2020). Accessed at: https://www.nist.gov/system/files/documents/2020/01/16/NIST%20Privacy%20Framework_V1.0.pdf

17. PHUSE: A Global View of the Clinical Transparency Landscape – Best Practices Guide, Version 1.0. (22 May 2020). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Clinical+Trials+Data+Transparency+Toolkit+Best+Practices+Guide.pdf

18. PHUSE: De-Identification Standard for CDISC SDTM 3.2, Version 1.01. (20 May 2015). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+Standard+for+SDTM+3.2+Version+1.0.xls

19. El Emam, K. (2013). Guide to the De-Identification of Personal Health Information. Auerbach Publications.

20. Information and Privacy Commissioner of Ontario: De-identification Guidelines for Structured Data. (June 2016). Accessed at: https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

21. European Medicines Agency: External guidance on the implementation of the European Medicines Agency Policy 0070 on the publication of clinical data for medicinal products for human use, Version 1.5. (14 May 2025). Accessed at: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use-version-15_en.pdf

22. El Emam, K., & Arbuckle, L. (2013). Anonymizing Health Data. O'Reilly.

23. Article 29 Data Protection Working Party: Opinion 05/2014 on Anonymisation Techniques, WP216. (10 April 2014). Accessed at: https://iapp.org/media/pdf/resource_center/wp216_Anonymisation-Techniques_04-2014.pdf

24. International Organization for Standardization: ISO/IEC 20889:2018(en) Privacy enhancing data de-identification terminology and classification of techniques. (November 2018). Accessed at: https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en:term:3.32

25. Health Canada: Public Release of Clinical Information, Version 1.0. (12 March 2019). Accessed at: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html

26. United States Department of Justice. Freedom of Information Act Improvement Act of 2016. 5 USC § 552. (30 June 2016). Accessed at: https://www.congress.gov/114/plaws/publ185/PLAW-114publ185.pdf

27. European Medicines Agency: European Medicines Agency policy on access to documents (POLICY/0043). (4 October 2018). Accessed at: https://www.ema.europa.eu/en/documents/other/policy-43-european-medicines-agency-policy-access-documents_en.pdf

28. Institute of Medicine (US). (2015). Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. National Academies Press.

29. The Government of Japan. Act on Access to Information Held by Administrative Organs. Act No. 42 of 1999. (14 May 1999). Accessed at https://www.soumu.go.jp/english/gyoukan/engv1_03.pdf

30. Pharmaceuticals and Medical Devices Agency. Accessed at: https://www.pmda.go.jp/english/index.html (last accessed 20 December 2022)

31. European Medicines Agency: Clinical Data Publication Portal. Accessed at: https://clinicaldata.ema.europa.eu/web/cdp/home (last accessed 20 December 2022)

32. European Medicines Agency: EU Clinical Trials Regulation (EU CTR) No. 536/2014. (16 April 2014). Accessed at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014R0536

33. European Medicines Agency: Clinical Trials Information System. Accessed at: https://euclinicaltrials.eu/search-for-clinical-trials/?lang=en (last accessed 20 December 2022)

34. Health Canada: Clinical Information Portal. Accessed at: https://clinical-information.canada.ca/search/ci-rc (last accessed 20 December 2022)

35. Vivli. Accessed at: https://vivli.org/ (last accessed 20 December 2022)

36. ClinicalStudyDataRequest.com. Accessed at: https://clinicalstudydatarequest.com/ (last accessed 20 December 2022)

37. The YODA Project. Accessed at: https://yoda.yale.edu/ (last accessed 20 December 2022)

38. The YODA Project: Community Data Sharing Resources. Accessed at: https://yoda.yale.edu/about/community-data-sharing-resources (last accessed 20 December 2022)

39. Clinical Data Interchange Standards Consortium: SDTM, V2.0. (29 November 2021). Accessed at: https://www.cdisc.org/standards/foundational/sdtm/sdtm-v2-0

40. Clinical Data Interchange Standards Consortium: ADaM, V2.1. (7 December 2009). Accessed at: https://www.cdisc.org/standards/foundational/adam/adam-v2-1

41. International Council for Harmonisation: Structure and Content of Clinical Study Reports E3. (30 November 1995). Accessed at: https://database.ich.org/sites/default/files/E3_Guideline.pdf

42. International Organization for Standardization: ISO/IEC 27559:2022 Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework. (November 2022). Accessed at: https://www.iso.org/standard/71677.html

43. Arbuckle. A. (2020). "Aggregated data provides a false sense of security." International Association of Privacy Professionals. Accessed at: https://iapp.org/news/a/aggregated-data-provides-a-false-sense-of-security

44. PHUSE Educate the General Population Project: Data Privacy and Data Sharing in Clinical Trials. Accessed at: https://advance.hub.phuse.global/wiki/spaces/WEL/pages/26805754/Educate+the+General+Population+on+Data+Privacy+and+Data+Sharing (last accessed 10 May 2025)

45. Bamford, S. et al. (2022). "Sharing Anonymized and Functionally Effective (SAFE) Data Standard for Safely Sharing Rich Clinical Trial Data." Applied Clinical Trials 31, 7/8. Accessed at: https://www.appliedclinicaltrialsonline.com/view/sharing-anonymized-and-functionally-effective-safe-data-standard-for-safely-sharing-rich-clinical-trial-data (last accessed 02 July 2023)

46. TransCelerate Biopharma Inc.: Clinical Data Sharing: A Proposed Methodology to Enable Data Privacy While Improving Secondary Use (August 2023). Accessed at: https://www.transceleratebiopharmainc.com/wp-content/uploads/2023/08/FINAL-Privacy-Methodology-Revision-August-25.pdf

47. Murdoch, B. (2021). "Privacy and artificial intelligence: challenges for protecting health information in a new era". BMC Medical Ethics 22:122. Accessed at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8442400/ (last accessed 28 September 2024)

7: Disclaimer

The opinions expressed in this document are those of the authors and should not be construed to represent the opinions of PHUSE members, respective companies/organisations or regulators' views or policies. The content in this document should not be interpreted as a data standard and/or information required by regulatory authorities.

8: Appendices

Appendix 1: Best Formatting Practices

Appendix 2: Examples of Structured IPD Anonymisation

Appendix 3: Examples of Unstructured IPD Anonymisation

Note: The examples presented in the appendices are not an exhaustive list and highlight the most frequently observed scenarios.

9: Project Contact Information

- Lauren Hepburn
- Abby McDonell
- Devaki Thavarajah
- Mahesh Swaminathan

Email: workinggroups@phuse.global

10: Acknowledgements

Mekhala Acharya, Asmita Anil, Sanjay Bagani, Sarah Balay, Cara Campora, Holly Curry, Dhiraj Ravjibhai Dabhi, Shalini Dwivedi, Rama Empati, Swagata Ghosh, Manohara Basoor Halasiddappa, Anitha Isaiah, Thomas Kalfas, Shirisha Kanthala, Michael McTaggart, Muhammad Oneeb Rehman Mian, Sharon Conder Niedecken, Pooja Phogat, Viveka Rydell-Anderson, Taniya Rade, Mithun Kumar Ranga, Benjamin C. Shim, Shweta Srivastava, Abhinav Srivastva, Tamsin Sargood, Brenda Tiffin, Simin Takidar, Thomas Wicks, Sophia Zilber.

The Working Group would like to thank representatives from both Health Canada and the European Medicines Agency for their support as observers in the creation of the Good Transparency Practice Guideline.

Appendix 1. Best Formatting Practices

1: Overview: Purpose of This Document

This appendix contains suggestions and examples of best formatting practices for both structured and unstructured individual patient data that the Data Controller should consider implementing to make the anonymisation process as efficient as possible.

2: Structured Individual Patient Data Formatting

- Follow SDTM/ADaM standards [1,2] for formatting datasets, ensuring adherence to naming conventions for tables and variables. If a dataset is not in SDTM formatting and a non-SDTM format must be used, then submit as much supporting documentation as possible.
- Create and retain documentation of how the ADaM datasets were created from the SDTM datasets.
- Anonymise SDTM and ADaM data together if possible.
- Variables should have detailed labels.
- Non-standard variables should be intuitively named.
- Imputed values in ADaM datasets should be clearly flagged.
- Coded variables should have a decode provided, either via a coded text response column (such as --STRESC and --STRESN), a metadata table, or a supporting document.
- Variables should be coded using industry standards, such as the Medical Dictionary for Regulatory Activities (MedDRA) [3] or the Anatomical Therapeutic Chemical (ATC) classification system [4].
- Standardise free-form text responses wherever possible (e.g. “Patient missed final treatment because they were unable to get transportation to site” could be standardised to “Patient discontinued treatment”).
- Avoid spreading information for a single variable across multiple columns when it could be compiled under one Key-Value pair of variables.
- Guidance on handling specific variables can be found in the PHUSE De-identification Standards [5].

3: Unstructured Individual Patient Data Formatting (These points may be enforced by style guides or template instructions)

The industry is moving towards taking a proactive approach from the beginning of clinical study report (CSR) development itself, by focusing on authoring the documentation in scope for public release to better facilitate anonymisation [6].

- Follow ICH E3 guidelines [7] for formatting clinical study reports, ensuring adherence to section nomenclature and numbering to allow for ease of out-of-scope content identification and removal. Avoid shifting the numbering of sections, keep sections even if not applicable, and, if adding sections, try and fit within ICH numbering or place them into appendices at the end of the document.
- Limit non-English language pages, but, if unavoidable, do not include PPD.
- Limit repetitive information. Instead, use cross-references within or between documents.
- Avoid using small text (e.g. font size 9+ recommended).
- Use staff names consistently throughout documents, limiting the variations used (e.g. use the format John Smith only rather than John Smith, J. Smith and Smith, John).
- Avoid using staff initials or usernames.
- Avoid using staff contact email addresses, telephone or fax numbers.
- Avoid exact titles. Use generic titles whenever possible.
- Avoid including staff names, initials, or usernames in table/figure captions or document headers/footers.
- Keep PPD in the main body of text (e.g. do not include PPD in the file name, header, footer, or bookmarks).
- Avoid including non-machine-readable PPD (e.g. do not include PPD embedded in figures, scanned pages and handwritten notes except for signatures).
- Avoid including patient-related images such as photographs or X-rays.
- Use a consistent subject ID format and labelling throughout the document for ease of pseudonymisation.
- Do not drop leading zeros in subject IDs (e.g. “PROTOCOL01-001001” and “001001” are both acceptable formats, while “1001” is not).
- Avoid very short subject IDs (less than 5 digits), as they are harder to distinguish from other numerical values such as dates or phone numbers.
- Avoid linking identifiers without a subject ID present (e.g. a 25-year-old man had an SAE of appendicitis).
- Be considerate when adding line breaks during the formatting stage (e.g. adding line breaks within a patient narrative).
- Consider a unified approach for publishing the documents (landscape vs portrait) depending on the information being presented (e.g. use portrait unless a table is wide with many columns, or a wide figure).

- Avoid splitting words, phrases or numbers across multiple lines in a document that will be rendered to PDF before use. For tables, consider adjusting column widths, condensing columns, using acceptable abbreviations, and/or rotating the page to landscape to allow for wider tables. (See the example below.)
    - Example of poor formatting:

Subject ID	Age	Sex	Race	Treatment	Treatment Start Date	Treatment End Date
US001-001001	26	Male	White	Placebo	2019-04-19	2020-05-02
US001-001002	36	Female	White	Treatment 1	2019-04-02	2019-07-03
US002-001003	20	Male	Multiple	Treatment 2	2019-03-22	2020-04-04
US003-001004	22	Female	Black or African American	Placebo	2019-03-19	2020-04-01

    - Example of good formatting:

Subject ID	Age/Sex/Race	Treatment	Treatment Start Date	Treatment End Date
US001-001001	26/M/White	Placebo	2019-04-19	2020-05-02
US001-001002	36/F/White	Treatment 1	2019-04-02	2019-07-03
US002-001003	20/M/Multiple	Treatment 2	2019-03-22	2020-04-04
US003-001004	22/F/Black or African American	Placebo	2019-03-19	2020-04-01

- Be considerate when creating aggregate tables in-text, when n=1 (or other small populations) for a subgroup, as a person's details could be inferred. (See the example below.)

	Subgroup 1 (n=5)	Subgroup 2 (n=1)
Average Age **(Min, Max)**	34 (18,60)	25 (25,25)

- Avoid unnecessarily grouping patients by shared sites, dates, medical histories or demographic subgroups, etc. (See the example below.)

Site ID	Subject ID
US001	001001 001002
US002	001003
US003	001004

- Avoid presenting information as checkboxes. (See the example below.)

Race	
White	
Asian	X
Black or African American	
American Indian or Alaska Native	
Native Hawaiian or Other Pacific Islander	
Multiple	
Other	

- Avoid presenting information as yes/no answers (e.g. “History of depression: Yes”).
- Avoid gendered pronouns. Instead, use “the patient” or “the participant”.
- Use International Organization for Standardization (ISO) [8] recommended units and abbreviations for all participants (e.g. three-letter country codes).
- Use MedDRA terminology [3] when describing adverse events and medical histories, even if it causes the sentence to read oddly.
- Consider including multiple MedDRA levels [3]. This can increase data utility, as the higher level can be retained while the lower is suppressed.
- Consider having medical histories in tables rather than written paragraphs (e.g. in prosaic narratives) so that replacement values cause less shifting.
- For concomitant medications, use the active substance name instead of the trade name when describing concomitant medications. Weigh the need for relevance of concomitant medications and whether the dose regimen adds value.
- Avoid verbal representation of typically numeric information (e.g. “two years old”).
- Use consistent phrasing whenever possible (e.g. 25 years old, 25-year-old, 25 years of age).
- Use clear and consistent date formats:
    - If dates can be avoided, use relative dates such as “Day 22”.
    - Limit the number of date formats to as few as possible (e.g. only use one numeric and one character date format).
    - Avoid ambiguous date formats (e.g. “07/05/12” could be interpreted as “07 May 2012” or “05 July 2012” or “12 May 2007”).
    - Avoid date formats that are region-specific (e.g. MM/DD/YYYY).
    - Avoid combining multiple dates in a single phrase (e.g. “1st to 5th of April 2015” could be written as “01 April 2015 to 05 April 2015”).
- Consider if the phrasing would allow an attacker to undo de-identification (e.g. “is a smoker”, “was a smoker”, “has never smoked” could be written as “Nicotine usage: Current smoker/Former smoker/Never smoked”).
- Limit details about the participant’s personal life unless medically relevant:
    - Avoid mentioning a participant’s marital status or family members unless describing a relevant family medical history or partner pregnancy (e.g. “His daughter called an ambulance” could be written as “An ambulance was called”).
    - Avoid mentioning a participant’s job or employment status. If it’s necessary to include it, avoid their job title (e.g. “The patient was diagnosed with depression after they were fired from their job” could be written as “The patient was diagnosed with depression after a personal hardship”).
    - Avoid giving details of the participant’s housing situation (e.g. “The patient fractured their wrist after falling down the front steps of their apartment complex” could be instead written as “The patient fractured their wrist after falling down a set of stairs”).
    - Avoid giving region-specific details of the participant’s location (e.g. “The patient discontinued treatment because they moved to a new state” could be written as “The patient discontinued treatment because they are unable to attend follow-up visits at site”).
    - Avoid releasing sensitive health information about a participant, unless it may be medically relevant to the case. Sensitive information may be identifying (e.g. a disability in a limb, an identifying scar, a culturally specific attribute) or stigmatising (e.g. sexual orientation, HIV infection status, outcome of pregnancy).
- Format summaries (protocol, results and plain language) using existing guidelines:
    - Centers for Disease Control and Prevention (CDC) citation/link [9]
    - ClinicalTrials.gov (CT.gov) plain language checklist citation/link [10]
    - EFPIA lay language principles citation/link [11]
    - European Commission (EC) citation/link [12]
    - European Forum for Good Clinical Practice (EFGCP) citation/link [13]
    - Multi-Regional Clinical Trials (MRCTs) citation/link [14]
    - UK Health Research Authority (HRA) guidance on plain language writing citation/link [15]

Note: Since the landscape is subject to change, please refer to existing regulatory guidance and resources for best practices on the preparations of these items summarised above.

4: References

1. Clinical Data Interchange Standards Consortium: SDTM, v2.0 (29 November 2021). Accessed at: https://www.cdisc.org/standards/foundational/sdtm/sdtm-v2-0

2. Clinical Data Interchange Standards Consortium: ADaM, v2.1 (7 December 2009). Accessed at: https://www.cdisc.org/standards/foundational/adam/adam-v2-1

3. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use: Medical Dictionary for Regulatory Activities. Accessed at: https://www.meddra.org

4. World Health Organization Collaborating Centre for Drug Statistics Methodology: Anatomical Therapeutic Chemical (ATC) Classification System. Accessed at: https://www.whocc.no/atc_ddd_index/

5. PHUSE: De-Identification Standard for CDISC SDTM 3.2, Version 1.01. (20 May 2015). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+Standard+for+SDTM+3.2+Version+1.0.xls

6. PHUSE Data Transparency Autumn Event 2024. Writing Better CSRs to Facilitate Anonymization for CDP/PRCI/CTIS. Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Data+Transparency/EU/Virtual+%E2%80%93+Autumn+Meeting/PRE_DT11.pdf

7. International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use: Structure and Content of Clinical Study Reports E3. (30 November 1995). Accessed at: https://database.ich.org/sites/default/files/E3_Guideline.pdf

8. International Organization for Standardization: ISO 3166 Country Codes. Accessed at: https://www.iso.org/iso-3166-country-codes.html

9. Centers for Disease Control and Prevention (CDC). Plain Language Materials & Resources. Accessed at: https://www.cdc.gov/health-literacy/php/develop-materials/plain-language.html?CDC_AAref_Val=https://www.cdc.gov/healthliteracy/developmaterials/plainlanguage.html (last accessed 10 May 2025)

10. ClinicalTrials.gov (CT.gov). Plain Language Guide to Write a Brief Summary. Accessed at: https://clinicaltrials.gov/submit-studies/prs-help/plain-language-guide-write-brief-summary (last accessed 10 May 2025)

11. European Federation of Pharmaceutical Industries and Associations (EFPIA). REFLECTION PAPER - EFPIA Guiding Principles on Layperson Summary. Accessed at: https://www.efpia.eu/media/25661/reflection-paper-efpia-guiding-principles-on-layperson-summary.pdf (last accessed 26 April 2023)

12. European Commission (EC). (2021). Good Lay Summary Practice. Accessed at: https://health.ec.europa.eu/system/files/2021-10/glsp_en_0.pdf (last accessed 26 April 2023)

13. European Forum for Good Clinical Practice (EFGCP). Accessed at: https://efgcp.eu/ (last accessed 26 April 2023)

14. Multi-Regional Clinical Trials (MRCT). Health Literacy in Clinical Research. Accessed at: https://mrctcenter.org/health-literacy/ (last accessed 26 April 2023)

15. UK Health Research Authority (HRA). Writing a plain language (lay) summary of your research findings. Accessed at: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/writing-plain-language-lay-summary-your-research-findings/ (last accessed 26 April 2023)

5: Disclaimer

The opinions expressed in this document are those of the authors and should not be construed to represent the opinions of PHUSE members, respective companies/organisations or regulators' views or policies. The content in this document should not be interpreted as a data standard and/or information required by regulatory authorities.

6: Project Contact Information

- Lauren Hepburn
- Abby McDonell
- Devaki Thavarajah
- Mahesh Swaminathan

Email: workinggroups@phuse.global

7: Acknowledgements

Mekhala Acharya, Asmita Anil, Sanjay Bagani, Sarah Balay, Cara Campora, Holly Curry, Dhiraj Ravjibhai Dabhi, Shalini Dwivedi, Rama Empati, Swagata Ghosh, Manohara Basoor Halasiddappa, Anitha Isaiah, Thomas Kalfas, Shirisha Kanthala, Michael McTaggart, Muhammad Oneeb Rehman Mian, Sharon Conder Niedecken, Pooja Phogat, Viveka Rydell-Anderson, Taniya Rade, Mithun Kumar Ranga, Benjamin C. Shim, Shweta Srivastava, Abhinav Srivastva, Tamsin Sargood, Brenda Tiffin, Simin Takidar, Thomas Wicks, Sophia Zilber.

The Working Group would like to thank representatives from both Health Canada and the European Medicines Agency for their input in the creation of the Good Transparency Practice Guideline.

Appendix 2. Examples of Structured IPD Anonymisation

ORIGINAL DATASET

STUDYID	DOMAIN	USUBJID	COUNTRY	SITEID	SEX	BRTHDTC	AGE	RACE	RACOTH	ETHNIC	RFICDTC	DTHDTC	ARM
PROTOCOL01	DM	PROTOCOL01-US001-001001	USA	US001	M	1995-04-01	26	WHITE	NULL	HISPANIC OR LATINO	2019-04-19		PLACEBO
PROTOCOL01	DM	PROTOCOL01-US001-001002	USA	US001	F	1985-03-01	36	WHITE	NULL	NOT HISPANIC OR LATINO	2019-04-02		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-US002-001003	USA	US002	M	2001-04-05	20	MULTIPLE	MIXED RACE ASIAN/WHITE	NOT HISPANIC OR LATINO	2019-03-22		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-US003-001004	USA	US003	F	1999-04-05	22	BLACK OR AFRICAN AMERICAN	NULL	NOT HISPANIC OR LATINO	2019-03-19		PLACEBO
PROTOCOL01	DM	PROTOCOL01-CA001-001001	CAN	CA001	M	1990-01-23	31	WHITE	NULL	HISPANIC OR LATINO	2019-03-16		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-CA001-001002	CAN	CA001	F	1981-10-19	40	WHITE	NULL	NOT HISPANIC OR LATINO	2019-03-18	2019-09-05	TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-BE001-001001	BEL	BE001	M	1975-09-23	46	WHITE	NULL	NOT HISPANIC OR LATINO	2019-03-23		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-BE002-001002	BEL	BE002	F	1983-06-01	38	BLACK OR AFRICAN AMERICAN	NULL	NOT HISPANIC OR LATINO	2019-03-21		PLACEBO
PROTOCOL01	DM	PROTOCOL01-BE002-001003	BEL	BE002	M	1964-03-01	57	BLACK OR AFRICAN AMERICAN	NULL	NOT HISPANIC OR LATINO	2019-04-04		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-IT001-001001	ITA	IT001	M	1995-11-23	26	WHITE	NULL	NOT HISPANIC OR LATINO	2019-03-23	2019-07-25	TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-IT001-001002	ITA	IT001	F	1992-04-05	29	BLACK OR AFRICAN AMERICAN	NULL	NOT HISPANIC OR LATINO	2019-04-02		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-IT002-001003	ITA	IT002	M	1986-05-07	35	WHITE	NULL	HISPANIC OR LATINO	2019-03-19		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-IT002-001004	ITA	IT002	F	1991-12-06	30	WHITE	NULL	HISPANIC OR LATINO	2019-03-25		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-BR001-001001	BRA	BR001	F	1979-12-15	42	WHITE	NULL	HISPANIC OR LATINO	2019-03-19	2019-09-23	TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-BR002-001002	BRA	BR002	M	2001-05-01	20	BLACK OR AFRICAN AMERICAN	NULL	HISPANIC OR LATINO	2019-04-03		PLACEBO
PROTOCOL01	DM	PROTOCOL01-BR002-001003	BRA	BR002	F	1985-08-14	36	BLACK OR AFRICAN AMERICAN	NULL	NOT HISPANIC OR LATINO	2019-04-04		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-AR001-001001	ARG	AR001	M	1963-12-29	58	WHITE	NULL	HISPANIC OR LATINO	2019-03-17		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-AR001-001002	ARG	AR001	F	1972-08-31	49	WHITE	NULL	HISPANIC OR LATINO	2019-03-19		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-AL001-001001	ALG	AL001	M	1975-06-17	46	OTHER	MIDDLE EASTERN	HISPANIC OR LATINO	2019-03-25		PLACEBO
PROTOCOL01	DM	PROTOCOL01-AL002-001002	ALG	AL002	F	1988-06-05	33	BLACK OR AFRICAN AMERICAN	NULL	HISPANIC OR LATINO	2019-03-24		TREATMENT 1

TRANSFORMED DATASET 1 (more controlled disclosure context)

STUDYID	DOMAIN	USUBJID	COUNTRY	SITEID	SEX	BRTHDTC	AGE	RACE	RACOTH	ETHNIC	RFICDTC	DTHDTC	ARM
PROTOCOL01	DM	PROTOCOL01-FP195-843195	North America	NULL	M	NULL	(25–29)	WHITE	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-ZR310-066432	North America	NULL	F	NULL	(35–39)	WHITE	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-GP023-332941	North America	NULL	M	NULL	(20–24)	NULL	NULL	NULL	2019-03-15		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-MK329-278395	North America	NULL	F	NULL	(20–24)	BLACK OR AFRICAN AMERICAN	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-PE082-382816	North America	NULL	M	NULL	(30–34)	WHITE	NULL	NULL	2019-03-15		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-OU420-128401	North America	NULL	F	NULL	(40–44)	WHITE	NULL	NULL	2019-03-15	2019-09-02	TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-HI235-068199	Europe	NULL	M	NULL	(45–49)	WHITE	NULL	NULL	2019-03-15		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-HP392-908493	Europe	NULL	F	NULL	(35–39)	BLACK OR AFRICAN AMERICAN	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-VD491-257351	Europe	NULL	M	NULL	(55–59)	BLACK OR AFRICAN AMERICAN	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-NK129-348305	Europe	NULL	M	NULL	(25–29)	WHITE	NULL	NULL	2019-03-15	2019-07-17	TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-YY410-237533	Europe	NULL	F	NULL	(25–29)	BLACK OR AFRICAN AMERICAN	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-IA477-853814	Europe	NULL	M	NULL	(35–39)	WHITE	NULL	NULL	2019-03-15		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-JO148-229530	Europe	NULL	F	NULL	(30–34)	WHITE	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-HU414-172345	South America	NULL	F	NULL	(40–44)	WHITE	NULL	NULL	2019-03-15	2019-09-19	TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-BB791-609324	South America	NULL	M	NULL	(20–24)	BLACK OR AFRICAN AMERICAN	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-KL229-385935	South America	NULL	F	NULL	(35–39)	BLACK OR AFRICAN AMERICAN	NULL	NULL	2019-03-15		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-ZD079-619304	South America	NULL	M	NULL	(55–59)	WHITE	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-UD409-359322	South America	NULL	F	NULL	(45–49)	WHITE	NULL	NULL	2019-03-15		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-DA910-072536	Africa	NULL	M	NULL	(45–49)	NULL	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-LS591-282395	Africa	NULL	F	NULL	(30–34)	BLACK OR AFRICAN AMERICAN	NULL	NULL	2019-03-15		TREATMENT 1

KEY
Pseudonymisation: Masking of the unique subject ID with an encrypted value that has the same length as the original ID
Generalisation: Reducing the precision of the field
Suppression: Replacing the original value with an empty cell
Date-shifting: Offsetting a date value according to the scheme defined in the PHUSE CDISC SDTM anonymisation standard [1]. This scheme determines a delta for each patient based on a difference between a date in the trial available for all patients (in this case, the first visit date/RFICDTC) and an anchor date (in this case, 15 March 2019).
Retain: Maintaining the original values
***NOTE: De-identified datasets may also shuffle the rows in a table to reduce clustering by site, start date, etc. This was not done in this example, to make the transformations easier to compare to the original.

TRANSFORMED DATASET 2 (less controlled disclosure context)

STUDYID	DOMAIN	USUBJID	COUNTRY	SITEID	SEX	BRTHDTC	AGE	RACE	RACOTH	ETHNIC	RFICDTC	DTHDTC	ARM
PROTOCOL01	DM	PROTOCOL01-FP195-843195	Rest of World	NULL	M	NULL	(20–29)	NULL	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-ZR310-066432	Rest of World	NULL	F	NULL	(30–39)	NULL	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-GP023-332941	Rest of World	NULL	M	NULL	(20–29)	NULL	NULL	NULL	2019-03-15		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-MK329-278395	Rest of World	NULL	F	NULL	(20–29)	NULL	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-PE082-382816	Rest of World	NULL	M	NULL	(30–39)	NULL	NULL	NULL	2019-03-15		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-OU420-128401	Rest of World	NULL	F	NULL	(40–49)	NULL	NULL	NULL	2019-03-15	2019-09-02	TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-HI235-068199	Europe	NULL	M	NULL	(40–49)	NULL	NULL	NULL	2019-03-15		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-HP392-908493	Europe	NULL	F	NULL	(30–39)	NULL	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-VD491-257351	Europe	NULL	M	NULL	(50–59)	NULL	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-NK129-348305	Europe	NULL	M	NULL	(20–29)	NULL	NULL	NULL	2019-03-15	2019-07-17	TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-YY410-237533	Europe	NULL	F	NULL	(20–29)	NULL	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-IA477-853814	Europe	NULL	M	NULL	(30–39)	NULL	NULL	NULL	2019-03-15		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-JO148-229530	Europe	NULL	F	NULL	(30–39)	NULL	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-HU414-172345	Rest of World	NULL	F	NULL	(40–49)	NULL	NULL	NULL	2019-03-15	2019-09-19	TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-BB791-609324	Rest of World	NULL	M	NULL	(20–29)	NULL	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-KL229-385935	Rest of World	NULL	F	NULL	(30–39)	NULL	NULL	NULL	2019-03-15		SCREEN FAILURE
PROTOCOL01	DM	PROTOCOL01-ZD079-619304	Rest of World	NULL	M	NULL	(50–59)	NULL	NULL	NULL	2019-03-15		TREATMENT 1
PROTOCOL01	DM	PROTOCOL01-UD409-359322	Rest of World	NULL	F	NULL	(40–49)	NULL	NULL	NULL	2019-03-15		TREATMENT 2
PROTOCOL01	DM	PROTOCOL01-DA910-072536	Rest of World	NULL	M	NULL	(40–49)	NULL	NULL	NULL	2019-03-15		PLACEBO
PROTOCOL01	DM	PROTOCOL01-LS591-282395	Rest of World	NULL	F	NULL	(30–39)	NULL	NULL	NULL	2019-03-15		TREATMENT 1

KEY
Pseudonymisation: Masking of the unique subject ID with an encrypted value that has the same length as the original ID
Generalisation: Reducing the precision of the field
Suppression: Replacing the original value with an empty cell
Date-shifting: Offsetting a date value according to the scheme defined in the PHUSE CDISC SDTM anonymisation standard [1]. This scheme determines a delta for each patient based on a difference between a date in the trial available for all patients (in this case, the first visit date/RFICDTC) and an anchor date (in this case, 15 March 2019).
Retain: Maintaining the original values
***NOTE: De-identified datasets may also shuffle the rows in a table to reduce clustering by site, start date, etc. This was not done in this example, to make the transformations easier to compare to the original.

References

1. PHUSE: De-Identification Standard for CDISC SDTM 3.2, Version 1.01. (20 May 2015). Accessed at: https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+Standard+for+SDTM+3.2+Version+1.0.xls

Project Contact Information

- Lauren Hepburn
- Abby McDonell
- Devaki Thavarajah
- Mahesh Swaminathan

Email: workinggroups@phuse.global

Acknowledgements

Mekhala Acharya, Asmita Anil, Sanjay Bagani, Sarah Balay, Cara Campora, Holly Curry, Dhiraj Ravjibhai Dabhi, Shalini Dwivedi, Rama Empati, Swagata Ghosh, Manohara Basoor Halasiddappa, Anitha Isaiah, Thomas Kalfas, Shirisha Kanthala, Michael McTaggart, Muhammad Oneeb Rehman Mian, Sharon Conder Niedecken, Pooja Phogat, Viveka Rydell-Anderson, Taniya Rade, Mithun Kumar Ranga, Benjamin C. Shim, Shweta Srivastava, Abhinav Srivastva, Tamsin Sargood, Brenda Tiffin, Simin Takidar, Thomas Wicks, Sophia Zilber.

The Working Group would like to thank representatives from both Health Canada and the European Medicines Agency for their input in the creation of the Good Transparency Practice Guideline.

Appendix 3. Examples of Unstructured IPD Anonymisation

1: Overview: Purpose of This Document

This appendix contains additional unstructured IPD anonymisation examples, continuing from the Anonymisation Strategy 1 example provided on page 10 of the Data Recipient section.

2: Anonymisation Strategies

Example: Redaction, Generalisation, Retention

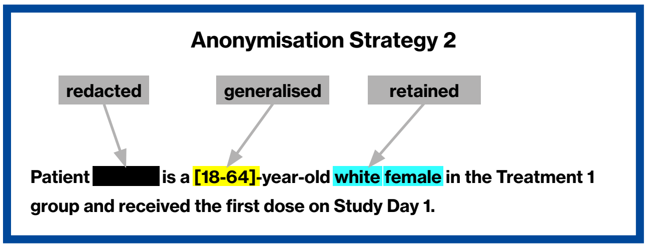

Anonymisation Strategy 2
redacted
generalised
retained
Patient [redacted] is a [18-64]-year-old white female in the Treatment 1 group and received the first dose on Study Day 1.

Figure 1. This example uses four different anonymisation techniques. The subject ID has been redacted, the age has been generalised, and the race and the gender have been retained. This strategy has retained most of the data utility.

Example: Redaction, Retention

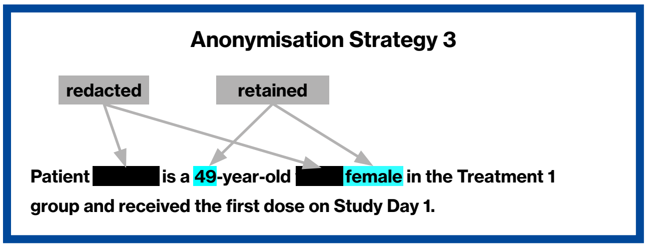

Anonymisation Strategy 3
redacted
retained
Patient [redacted] is a 49-year-old [redacted] female in the Treatment 1 group and received the first dose on Study Day 1.

Figure 2. In this example, the subject ID and the race have been redacted. The rest of the identifiers have been retained. This strategy has retained some data utility.

Example: Redaction

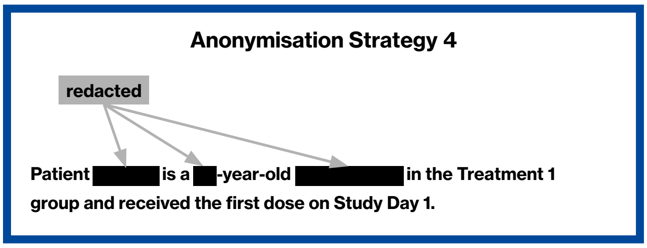

Anonymisation Strategy 4
redacted
Patient [redacted] is a [redacted]-year-old [redacted] in the Treatment 1 group and received the first dose on Study Day 1.

Figure 3. This example uses a targeted redaction strategy to continue protecting patient privacy when generalisation, retention and psuedonymisation cannot be used. This strategy has retained minimal data utility. Regulators prefer redactions to be targeted to only cover identifiers and will require the Data Controller to fill a deviation if blanket redactions are used to cover large portions of the text. Blanket redactions should only be considered as a last resort in extreme cases where low risk is not otherwise possible, and this decision needs to be robustly justified in the anonymisation report.

Original Narrative

Subject PROTOCOL01-AR001-001002

Reason for inclusion in narrative: AE (Tooth Abscess)

Preferred Term (Reported Term)	Onset Date (Study Day)	Date Resolved (Study Day)	Severity	Relationship	Action Taken	Outcome
Tooth Abscess (Dental Abscess)	2019-05-11 (54)	2019-05-14 (57)	SEVERE	NOT RELATED	Dose not changed	RESOLVED

Medical History

System Order Class	Preferred Term	Start Date	End Date
Infections - pathogen unspecified	Gingivitis	2019-03	2019-03
Musculoskeletal and connective tissue disorders	Chronic Pain	2015	Ongoing
Nervous system disorders	Migraine	1999	Ongoing

Baseline Characteristics

Site ID: AR001	**Age (yrs):** 49	**Treatment Group:** 2
Country: ARG	**Height (cm):** 172.5	**Study Start Date:** 2019-03-19
Race: White	**Weight (kg):** 80.3	**Study End Date:** 2020-03-30
Sex: F	**BMI (kg/m2):** 27.0	**Date of Death:** -

Narrative text

Patient 001002 is a 49-year-old white female in the Treatment 2 group and received the first dose on Study Day 1. The patient has a relevant medical history of gingivitis, chronic pain and migraines. No relevant concomitant medications were reported at baseline.

The last dose of the study drug before the adverse event was Study Day 49. On Study Day 54, the patient reported moderate jaw pain on the lower left side. The pain intensity progressed to severe on Study Day 56. The patient reported facial swelling and fever (39°C) on the same day. On Study Day 57, the patient presented themselves to the emergency room. An X-ray confirmed a periapical abscess of the lower left 1st molar. The patient underwent an emergency root canal. Ibuprofen and amoxicillin were prescribed to the patient. The patient was discharged from the hospital the same day. The adverse event of dental abscess was reported as resolved on Study Day 57. The investigator considered the adverse event unrelated to the study medication.

Anonymised Narrative (Transformation examples for narratives)

Subject **PROTOCOL01- UD409-359322**

Reason or inclusion in narrative: AE (Tooth Abscess)

Preferred Term (Reported Term)	Onset Date (Study Day)	Date Resolved (Study Day)	Severity	Relationship	Action Taken	Outcome
Tooth Abscess (Dental Abscess)	2019-05-07 (54)	2019-05-10 (57)	SEVERE	NOT RELATED	Dose not changed	RESOLVED

Medical History

System Order Class	Preferred Term	Start Date	End Date
Infections - pathogen unspecified	Gingivitis	[Date]	[Date]
Musculoskeletal and connective tissue disorders	Chronic Pain	[Date]	Ongoing
Nervous system disorders	Nervous system disorders	[Date]	Ongoing

Baseline Characteristics

Site ID: UD409	**Age (yrs):** 40-59	**Treatment Group:** 2
Country: South America	**Height (cm):** [**]	**Study Start Date:** 2019-03-15
Race: [Race]	**Weight (kg):** <90	**Study End Date:** 2020-03-26
Sex: F	**BMI (kg/m2):** [**]	**Date of Death:** -

Narrative text

Patient 359322 is a 40–59-year-old [Race] female in the Treatment 2 group and received the first dose on Study Day 1. The patient has a relevant medical history of gingivitis, chronic pain and nervous system disorders. No relevant concomitant medications were reported at baseline.

The last dose of the study drug before the adverse event was Study Day 49. On Study Day 54, the patient reported moderate jaw pain on the lower left side. The pain intensity progressed to severe on Study Day 56. The patient reported facial swelling and fever (39°C) on the same day. On Study Day 57, the patient presented themselves to the emergency room. An X-ray confirmed a periapical abscess of the lower left 1st molar. The patient underwent an emergency root canal. Ibuprofen and amoxicillin were prescribed to the patient. The patient was discharged from the hospital the same day. The adverse event of dental abscess was reported as resolved on Study Day 57. The investigator considered the adverse event unrelated to the study medication.

3: Disclaimer

The opinions expressed in this document are those of the authors and should not be construed to represent the opinions of PHUSE members, respective companies/organisations or regulators' views or policies. The content in this document should not be interpreted as a data standard and/or information required by regulatory authorities.

4: Project Contact Information

- Lauren Hepburn
- Abby McDonell
- Devaki Thavarajah
- Mahesh Swaminathan

Email: workinggroups@phuse.global

5: Acknowledgements

Mekhala Acharya, Asmita Anil, Sanjay Bagani, Sarah Balay, Cara Campora, Holly Curry, Dhiraj Ravjibhai Dabhi, Shalini Dwivedi, Rama Empati, Swagata Ghosh, Manohara Basoor Halasiddappa, Anitha Isaiah, Thomas Kalfas, Shirisha Kanthala, Michael McTaggart, Muhammad Oneeb Rehman Mian, Sharon Conder Niedecken, Pooja Phogat, Viveka Rydell-Anderson, Taniya Rade, Mithun Kumar Ranga, Benjamin C. Shim, Shweta Srivastava, Abhinav Srivastva, Tamsin Sargood, Brenda Tiffin, Simin Takidar, Thomas Wicks, Sophia Zilber.

The Working Group would like to thank representatives from both Health Canada and the European Medicines Agency for their input in the creation of the Good Transparency Practice Guideline.