Astellas Data De-Identification Standards

Table of Contents

1. Introduction

Providing access to data in ways that allows further research while maintaining the privacy and confidentiality of research participants is important for everyone involved in Astellas trials. There are several privacy law and regulatory guidance documents which need to be followed. This document describes the Astellas approach to prepare data for sharing with other researchers in a way that:

- Minimizes risks to the privacy and confidentiality of research participants.
- Ensures compliance with data privacy legal requirements.

2. General Approach

Upon approval of the research proposal by Independent Review Panel (IRP), the following data and relevant study documents are shared with the research team:

2.1 Raw study datasets
2.2 Analysis-ready datasets
2.3 Annotated Case Report Form (CRF)
2.4 Dataset specifications
2.5 Redacted Protocol with any amendments
2.6 Redacted Statistical Analysis Plan (SAP)
2.7 Redacted Clinical study report

Raw and analysis-ready datasets are de-identified by removing or replacing all Personally Identifiable Information (PII). Subject identifiers are recoded consistently across all datasets, to break any links with original study data or documentation, but ensure all data of one subject remains linked together. Free text fields (i.e. fields that contain data entered in the source database manually) are emptied. All dates are offset according to the algorithm described by the PhUSE organization [i].

3. Removing Personally Identifiable Information (PII) from the dataset

Identifiers as defined by HIPAA (see Code of Federal Regulations - Title 45: Public Welfare, Subtitle A §164.514, and related documentation) are not typically collected in study data, and are removed if present. In addition, any other PII that maybe present is removed or recoded. This involves:

- removing any names (of persons or institutions/companies) and initials,
- recoding kit numbers and device numbers
- Removing socioeconomic data including, but not limited to: occupation, income, education, household andfamily composition

Quasi-identifiers are also removed per PhUSE, some examples include:

- recoding identifiers (or code numbers) (see Section 3.1).
- removing free text verbatim terms (see Section 3.2).
- replacing date of birth with the categorized age (see Section 3.3).
- all original dates relating to individual subjects are shifted via an offset, and replaced with the new dummy dates so that the relative times between dates are retained (see Section 3.4).
- replacing country with a more-general geographic area (e.g., continent such as ‘North America’, region such as ‘Eastern Europe’).
- reviewing and removing/redacting other PII (see Section 3.5).

These steps are described in further detail below.

3.1 Recoding Identifiers (or code numbers)

The following identifiers are re-coded and the code key that was used to generate the new code numberfrom

the original code number is destroyed (as described in section 5):

- The investigator identifier (or code number) is re-coded or set to blank for each investigator. The investigator name is set to "blank".
- The new subject identifier for each research participant is consistently applied across all datasets in the study. The same new identifiers (or code numbers) are used across all datasets applicableto a single study e.g. raw dataset, analysis-ready dataset.
- Site identification information is re-coded or set to blank.

A proposal that includes multiple studies (e.g., extension studies, long term follow-up studies finished at the time of the receipt of the proposal) will use the same new identifiers as re-coded for the initial study to enable individual subject data to remain linked. This is achieved by repeating the data anonymization process for the initial study data at the same time as the extension/follow up data. However, an extension study for a separate proposal (e.g., that was not handled at the same time of the other proposal) will follow an independent data anonymization process.

3.2 Removing Free Text Verbatim Terms

Information in a descriptive free-text verbatim term may compromise a subject's anonymity. Free text verbatim terms are set to "blank" including:

- adverse events
- medications
- medical history
- other specific verbatim free text
- information from variable names e.g. lab names may contain location information
- investigator comments that may directly or indirectly be used to identify a subject
- genetic data that may enable a direct trace back to an individual subject

Certain free text fields may be retained or partially masked (see Section 3.5) if they do not contain PII and removal of these fields may impact the scientific value of the dataset (e.g. medical history that has not been coded).

In cases when a dataset only includes variables which are set to missing by the above process (e.g., a Subject Characteristic dataset containing only text fields which have been set to missing), the dataset can be dropped from de-identified database.

3.3 De-Identifying Date of Birth and Age

Information relating to a research participant's date of birth and identification of specific ages above 89 may compromise anonymity. The following steps will be taken:

- Date of birth is set to missing.
- In addition, the aggregate age as 5-year categories will be defined as a default (20-24,25-29, 30-39, etc.), however the age categorizations will be adjusted based on the number of the subject per category. All age categorizations will define ages above 89 into the category of ">89 years".
- Pediatric studies where age is collected in months or days will use the age categories recommended by PhUSE.
- By default, the continuous age variable will be dropped. If the inclusion of age (as a continuous variable) is necessary for a proposal's analysis, the continuous age variable can be retained (with values above 89 set to missing).
- See Section 4 on handling of age categories to reach the appropriate risk de-identification threshold.

3.4 Replacing all Original Dates relating to a Research Participant

Dates are offset according to the scheme defined by PhUSE. This scheme determines an offset for each participant based on a difference between a date in the trial available for all participants and an 'anchor date' (which is typically study initiation date).

For each research participant, the offset is applied to all dates. All original dates are replaced with the new dummy dates so that the relative times between dates are retained.

If an original date contains partial information, the following steps will be performed:

- an algorithm to impute the missing portions will be applied to make the variable complete,
- the offset is applied to the complete date,
- finally, the offset date will be displayed consistent with original partial date (e.g., an original date of "2020-06" might be displayed as "2019-11").

3.5 Reviewing and Removing/Redacting Other PII

Other data elements that contain PII could be redacted, if necessary. A free-text variable required for analysis (and not coded into another variable supported bycontrolled terminology) must be reviewed. Values with personal information within the string replaced with text to indicate that a value has been redacted. Example: "Dr Adam assessed tumor on right arm" becomes "Dr -XX- assessed tumor on right arm".

4. Calculation of Risk Re-Identification

The risk of Re-Identification will be calculated using the following quasi-identifiers: categorized age, sex, continent/region, race, and ethnicity. Variables will only be included if the maximum re-identification risk is below the following:

- CSDR Research Proposal or similar: 0.34 (e.g., there are at least 3 subjects with the same set of quasi-identifiers).
- Data available to public: 0.091 (e.g., there are at least 11 subjects with the same set of quasi-identifiers).

If needed, the categorization scheme will be updated to allow for large enough sample sizes (for example, all age groups >=70 years old might be combined). Quasi-identifiers which are unable to be combined to reach the appropriate risk threshold will be dropped from the datasets or set to missing.

5. Review and Quality Control

A final review of the assigned DI (de-identification) rule assignments is made to determine if further removal is required. Quality Control (QC) checks and documentation (QC record) is conducted for the processing of the data and supportive metadata documentation. This review will include:

- Confirming that the number of records in a dataset remains constant with the original dataset. If not, the reason must be investigated and explained, if it was done on purpose for de-identification.
- Checking that all specified changes were made to the datasets (For example, verifying that all date variables have been offset, and that text fields have been set to missing.)
- Verifying that no fields were changed, except as specified above.

6. Destroying the link (key code) between the dataset that is provided and the original dataset

Research participants' identification code numbers are de-identified by replacing the original code number

with a new code number (as described in 3.1) and destroying the code key that was used to generate the new code number from the original (i.e. destroying the link between the two code numbers).

The following specific items are discarded:

- any transactional copies of anonymized datasets
- de-identification tables (links from original variable to new anonymized variable)
- QC output datasets and review files
- Any SAS log or output (e.g. lst) files that contain PII

The de-identified datasets are stored in a separate secure location from the original datasets.

References:

[i] *Data Anonymisation and Risk Assessment Automation*, PHUSE Data Transparency Working Group. https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+Anonymisation+and+Risk+Assessment+Automation.pdf