Astellas Data Anonymization 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 laws and regulatory guidance documents which need to be followed (for example guidance from European data protection regulators and Code of Federal Regulations - Title 45: Public Welfare, Subtitle A §164.514). In addition, international working groups and publications in this area provide guidance [1], [2] to the CTDD (Clinical Trial Data Disclosure) team involved in the data anonymization process at Astellas.

This document describes 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 Protocol with any amendments
2.6 Statistical Analysis Plan (SAP)
2.7 Redacted Clinical study report

Raw and analysis-ready datasets are anonymized 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 removed as well; adding a study day variable, if that is not already present.

3. Removing personally identifiable information (PII) from the dataset

All identifiers as defined by HIPAA (see Code of Federal Regulations - Title 45: Public Welfare, Subtitle A §164.514, and related documentation) are removed from the datasets. In addition any other PII that may be present is removed.

This involves removing:

- any names (of persons or institutions/companies) and initials,
- (or recoding) kit numbers and device numbers

- de-identifying geographic information that applies to specific subject groups (< 10 subjects, e.g. site name, place of work). In some cases, country can be removed and replaced with continent.

- socioeconomic data including, but not limited to: occupation, income, education, household and family composition

In addition the following steps are undertaken:

- recoding identifiers (or code numbers) (see Section 3.1).

- removing free text verbatim terms (see Section 3.2).

- replacing date of birth with both age and the categorized age(see Section 3.3).

- all original dates relating to individual subjects are set to blank, providing the study day instead (see Section 3.4).

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

- The investigator identifier (or code number) is re-coded 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 applicable to 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

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).

3.3 Replacing Date of Birth

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 replaced with the age at reference start date. Ages above 89 are set to blank. If age is derived from a partial date, any flag variables in the analysis-ready datasets will also be kept.
- In addition, the aggregate age as 5-year categories will be defined as a default (20-24, 25-29, 30-39, etc.), while other age categorizations can be included if appropriate. All age categorizations will define ages above 89 into the category of ">89 years".
- If the inclusion of age (as a continuous variable) could compromise the participant anonymity, the age variable can be dropped any only the variable containing the aggregate age category would be retained.

3.4 Replacing all Original Dates relating to a Research Participant

Study Day Method:

All dates are set to blank. If not available in the dataset already, the study day is calculated for each observation with days relative to a reference date. If clearly defined in the dataset model or specifications, the reference date is used as defined therein. If not, in order of priority the reference date is defined as the date of first study treatment, date of randomization or date of consent. For example if a patient is randomized, but does not take the study treatment (i.e. the date of first treatment is missing), the date of randomization is used as the reference date to calculate the study day for any assessments recorded.

Example If the original reference date was 01JAN2008 and the date of death was 01MAY2008, the date of death would be 122 expressed as study days.

3.5 Reviewing and Removing/Redacting Other PII

Other data elements that contain PII are removed. For example:

- information from variable names e.g. lab names may contain location information

- investigator comments that may be used to identify a subject

- A free-text variable required for analysis (and not coded into another variable supported by controlled terminology) must be reviewed. Values with personal information within the string

replaced with "--redacted--". Example: "Dr Adam assessed tumor on right arm" becomes "--redacted-- assessed tumor on right arm".
- genetic data that may enable a direct trace back to an individual subject

4. 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 removed, and that relative study day is included.)
- Verifying that no fields were changed, except as specified above.

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

Some data protection authorities in Europe suggest that the data can only be considered anonymized if personal information is removed (or redacted) and the subject code number cannot be linked to a research participant. Therefore, research participants' identification code numbers are anonymized by 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).

Research participants' identification code numbers are anonymized 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 anonymized datasets are stored in a separate secure location from the original datasets.

References:

1 Hrynaszkiewicz I, Norton ML, *et al.* Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ* 2010; **340**: c181.
2 De-identification of Clinical Trials Data Demystified. *Jack Shostak, Duke Clinical Research Institute (DCRI), Durham, NC http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf*