April 1 2022

Sumitomo Pharma Anonymization of Clinical Trial Datasets

This document summarizes the data anonymization steps performed by Sumitomo Pharma (SMP) prior to data release.

Introduction

Sumitomo Pharma (SMP) believes that providing qualified researchers with access to clinical trial data, information and results will contribute to the advancement of public health and will ultimately help improve the speed of pharmaceutical product development for the good of all. However, SMP also recognizes that while such access should be provided, individual subject privacy and confidentiality must be protected at the highest level.

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.

The information contained in this document represents the general standards and methods SMP employs for anonymizing clinical trial data. However, in some cases, alternative approaches to anonymization than those described here may be employed to assure that individual subject privacy and confidentiality are always assured.

Handling Specific Data Fields

1. All 18 items identified in HIPPA will be considered in de-identifying datasets considered for release (see Table 1). Except for dates, all 17 items will be removed from all datasets or documents prior to release in addition to other quasi-identifiers from medical history, knowable events such as death, gender, and race. For specific trials there may be more quasi-identifiers that would need to be considered for deduction.
2. General rules for handling dates:
    - Any date that can be used to identify an individual subject's age will be removed
    - Dates in any way referring to an individual's life milestones (e.g., date of birth, date of death) or their actions (e.g. enrollment date, visit date, test date) will be handled as described below

Table 1: The HIPAA Identifiers

HIPAA Items	HIPAA Items
Names	Account numbers
All geographic subdivisions smaller than a state	Certificate/license numbers
Telephone numbers	Vehicle identifiers and serial numbers, including license plate numbers
Facsimile numbers	Device identifiers and serial numbers
Electronic mail addresses	Web universal resource locators (URLs)
Social Security numbers	Internet protocol (IP) address numbers
Medical record numbers	Biometric identifiers, including fingerprints and voiceprints
Health plan beneficiary numbers	Full-face photographic images and any comparable images
	Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification

3. Birth date/Date of death/Other publicly available dates (i.e., life-milestone dates):
    - Remove birth date/date of death and replace with age/age at death. However, ages over 89 may be aggregated into a single category of age 90 or older
    - Age specific outliers will not be included in released datasets; these dates will be reported by ranges (e.g., 'X to Y' or 'XX or older')
    - Sponsor statisticians will decide what and how age ranges/groups shall be pooled in de-identified datasets

4. Other dates:
    - Replace all original dates related to individual subject with new/revised dates offset by a random number provided by the statistician; for partial dates, impute the partial dates to middle of the month, middle of the year before offsetting
    - The statistician will use an unique seed for each study and that seed will be used to generate a random offset number to be applied for that individual subject consistently across all datasets to be shared
    - The seed used to generate random numbers will not be disclosed at any time

5. The above approach will be applied to all elements of dates directly related to an individual subject, including admission date and discharge date

6. Geographic Information:
    - Most geographic information that may be used to identify an individual subject will be removed or modified to prevent re-identification

- The statisticians will decide if there will be any geographic information retained. However, any geographic information that is retained will not provide sufficient detail allowing the re-identification of individual subjects
- For example, large regional designators (e.g., North America, Eastern Europe) may be used as a geographic specifier

7. Subject IDs:
    - Replace all subject IDs with a new, randomly-generated ID

8. Investigator/Site ID:
    - Remove all investigator names
    - Generally, site/investigator IDs/numbers will be removed, set to blank, or re-coded. In the latter case, each site is given a new randomly generated identifier.
    - Trials containing sites with <10 subjects will require case-by-case attention to assure that anonymization is not compromised. Consideration will be given to aggregation of small-N sites into one or a small number of 'larger sites' or possibly dropping center as a parameter in the datasets.
    - However, if it is necessary to retain site IDs, then actual site IDs will be replaced by randomly generated site IDs, in a manner similar to subject
    - ID replacement.

9. Genetic data:
    - All genetic data will be removed

10. Other PII:
    - Information from variable names or label (e.g. lab names may contain location information) may be renamed on case by cases
    - Exploratory Biomarker data outside the primary and key secondary endpoints and laboratory data will be removed
    - Case narratives, documentation for adjudication and imaging data (e.g. x-rays, MRI scans) will be removed

11. Free text/verbatim fields:
    - Remove all free text/verbatim terms

12. All standard dictionary terms will be retained

13. Investigator comments will be removed as they may be used to identify a subject

14. Reference IDs:

- Remove all KIT IDs and Assessment reference IDs (e.g., ECG, LAB, PK etc.)

15. Remnants:
    - After anonymization, no information will be retained by the sponsor (‘remnants’) that would allow the possibility to recreate the original datasets from the anonymized data.