Clinical Trial Data Sharing – Anonymization Standards

July 2024

PURPOSE

The purpose of this document is to provide technical guidance for the anonymization of participant data from clinical trials that are applicable to this guideline.

This document prescribes a minimum set of steps required to protect the privacy of the participants in clinical trials by anonymizing their personal data prior to sharing for secondary analyses via a controlled, secure and restricted access environment in accordance with applicable laws and regulations. In addition, since clinical trials have varying designs and data characteristics, this document promotes trial-specific evaluation of all data fields to determine if further steps of data anonymization are taken. This trial specific approach is especially important for clinical trials in rare diseases, pediatric clinical trials and clinical trials with small sample sizes.

ANONYMIZATION OF STUDY PARTICIPANT LEVEL CLINICAL TRIAL DATA

1. COMBINATION OF RULE-BASED AND RISK-BASED APPROACH

The anonymization of study participant level clinical trial data (SDTM and ADaM) is conducted at the trial level by combining the rule-based and the risk-based approach. The initial rule-based approach is systematically applied for each standard data field, followed by the risk-based approach to reinforce the protection of privacy if needed. This ensures that personal data is protected by minimizing the risk of re- identification and at the same time maintaining data utility.

- Quantitative risk based approach

  The quantitative risk of re-identification is computed to control the risk and adjust the banding of continuous quasi-identifier variables.

  The risk is called internal risk, i.e., limited to study data. Implementation is done according to recommendations from different articles (Protecting Privacy Using k-Anonymity - PMC (nih.gov).

  The risk methodology is based on k-anonymity and uses the five quasi-identifiers WEIGHT, AGE, GENDER, ETHNIC ORIGIN and GEOGRAPHICAL LOCATION to calculate the quantitative risk. When BMI is not available in the study, then HEIGHT will be classified and added to quantitative risk.

  Please also refer to https://advance.phuse.global/pages/viewpage.action?pageId=10878987 for further information on the definition of k-anonymity.

- Thresholds
  - The disclosure Risk Probability of Re-identification must be below **threshold 0.09**. If the risk is above, then weight and/or age classes are regrouped, and isolated values of ethnic origins or geographical location are generalized, up to Risk Probability of Re-identification is below the threshold.
  - The minimum is set to k=2 except for the maximum of 5% of the population for which we allow uniqueness representation in the population on the 5 quasi identifiers. If the proportion of unique records is more than 5%, then weight and/or age classes are regrouped, and isolated values of ethnic origins or geographical location are generalized, up to proportion of unique records is below the threshold.

- Protection of sensitive information based on L-DIVERSITY

  Adverse Event (AE), Medical History (MH), and Concomitant Medications (CM) are considered as sensitive information. Even after quasi-identifiers are generalized, dropped, or altered, and k- anonymity is equal or greater than 3 (k>=3), the sensitive information AE, MH, and Concomitant Medications may still not offer sufficient variability/diversity if the L- DIVERSITY is below 3 (<3). In this case, a study participant has the characteristics since all study participants have those same characteristics and the content of the sensitive variable is replaced by '--REDACTED--'.

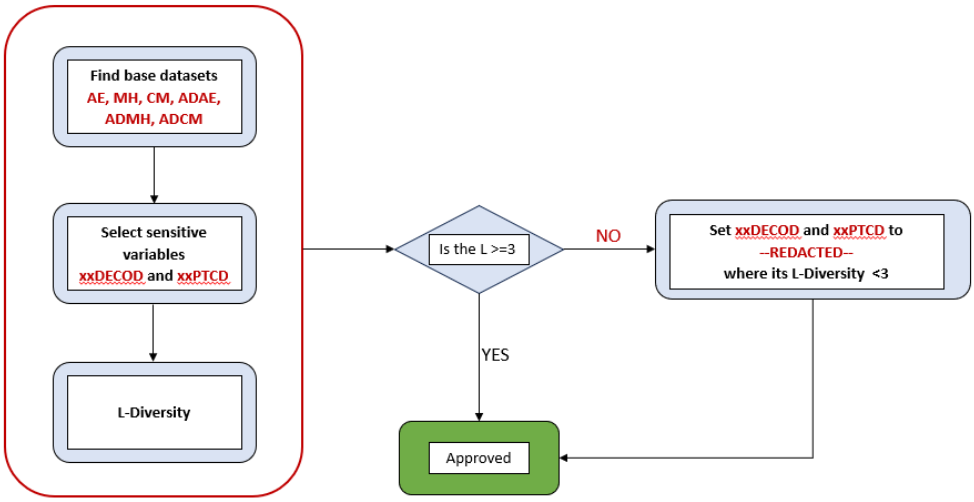

Figure 1: Sensitive information based on L-DIVERSITY

2. STANDARD DATA FIELDS CONSIDERED FOR ANONYMIZATION

All direct identifiers, as characterized by the Health Insurance Portability and Accountability Act (HIPAA)/Safe Harbor, are considered for removal, except for elements of dates and geographic information which are considered for transformation when anonymizing study participant level clinical trial data and related dataset documentation. This is to prevent the risk of association of a trial participant to his/her data.
Please refer to the Appendix for the list of the 18 HIPAA identifiers.

The following table of the standard data fields:

- Provides details of transforming the information related to dates and geographic information,
- Identifies additional attributes which are addressed when anonymizing data related to clinical trial participants.

The information in this table is not exhaustive. There are attributes which may be product, study phase and / or disease-specific and are addressed accordingly.

Data Field	By Default Anonymization Approach
HIPAA 18 Identifiers	● Dates and geographic information are altered according to the guidance below. ● All other 16 identifiers are removed. Yet they are mentioned as a reference but note that clinical trial data do not contain any of the 16 direct identifiers.
Study Participant Status	● Screen Failure study participants are dropped from datasets.
Study Participant ID Replacement	● Replace the original study participant ID by a new random study participant ID that cannot be linked back to the original study participant ID but can be still used to link information across domains within anonymized study database. ● Link key between actual initial study participant identifier and new random number is systematically deleted, which turns de-identification into anonymization.
Dates	● All dates are shifted. **DATE OFFSET** is a method where the original date is transformed by adding or subtracting a random number of days from the original date.
Birth date / AGE	● Unlike other dates, date of birth is removed and not shifted. ● Age is converted into fix classes.
Comments, FreeText, Verbatim	● Comments, Free text, and Verbatim fields replaced by "—REDACTED—".
Terms: **Adverse Event (AE),** **Concomitant Medication (CM),** **Medical History (MH)**	● Adverse Events: ○ Remove reported AE term by the investigator, only coded AE terms are kept, ○ Drop Lowest Level Term because too detailed. Coding hierarchy is kept starting from the Preferred Term on. ● Same for Medical History ● Concomitant Medication: only coded transcription in international name must be kept. ● Apply L-DIVERSITY to protect sensitive information
KIT Numbers & Device Numbers	● Remove kit numbers, device numbers, and other information linked to the treatment, such as lot numbers, batch numbers. ● Remove device numbers uniquely linked to study participants, e.g., pacemaker, to minimize risk of re- identification.

Data Field	By Default Anonymization Approach
Investigator ID & Name	● Remove investigator identifier.
Site ID	● Site ID is replaced in a similar mode as for Study Participant ID: replace the original site identifier by a new random site number that cannot be linked back to the original site ID, however, can be still used to link information within anonymized study datasets.
Geographic Information	● A predefined model of grouping countries is applied until all combinations of GENDER * ETHNIC ORIGIN * GEOGRAPHICAL LOCATION are >= 2 study participants. If a combination is still under 2, then group ETHIC ORIGIN (see below in Demographic Information block). ● Remove results in original units. Results in standard units are kept.
Demographic Information	● Hispanic ethnicity is removed. ● Aggregate ethnic origin with few study participants under “Unknown” ethnic origin group until all combinations of GENDER * ETHNIC ORIGIN * GEOGRAPHICAL LOCATION are >= 2 study participants.
Height	● HEIGHT is removed.
BMI	● BMI ○ For AGE >=20 years old, BMI is converted into following classes from WHO: Below 18.5 = Underweight; 18.5–24.9 = Normal weight; 25.0–29.9 = Pre-obesity; 30.0–34.9 = Obesity class I; 35.0–39.9 = Obesity class II; Above 40 = Obesity class III. ○ For pediatric studies, or mixed studies, please refer to World Health Organization (WHO) classification.
Weight	● Weight is converted into fix classes.
Genetic Data	● Remove all genetic data.
Interactive Voice Response System (IVRS), Randomization	● Remove IVRS and rando datasets

Data Field	By Default Anonymization Approach
Deviations	● Remove Deviations datasets
SUPPQUAL	● Systematically dropped, only “core” SDTM are kept.

Note that Trial Design Model is kept, as no personal data is part of it.

3. NON-STANDARD APPROACH CONSIDERED FOR ANONYMIZATION

In addition to the above standard data fields, any other indicator that could be used alone or in combination with other information to identify an individual who is subject of the information must be removed. This is part of the safe harbor method.

Clinical trials with small sample size and clinical trials in rare diseases

A more conservative approach is applied for studies with study participant numbers below 100 and may also be considered for studies in rare diseases. Quasi- identifiers such as GENDER, ETHNIC ORIGIN, BMI, WEIGHT, HEIGHT, and GEOGRAPHICAL LOCATION are removed in order to protect personal data.

Clinical trials with small duration

Date shifting will not necessarily be sufficient to prevent inference of certain dates in some cases. For instance, if a trial was run for less than a year, then the recipient of the data would have a bound for the date that is smaller than has been recommended by HIPAA Safe Harbor. Additional protections would need to be taken and could include replacing dates with relative study days.

4. ANONYMIZATION EXAMPLES

A. The following example shows a subset of a dataset before and after the data anonymization process by applying the minimum set of anonymization steps.

Before data anonymization:

USUBJID	SUBJID	SITEID	COUNTRY	AGE	SEX	RACE	RFSTDT	WGTBL	HGTBL	BMIBL
Unique Subject Identifier	**Subject Identifier for the Study**	**Study Site Identifier**	**Country**	**Age**	**Sex**	**Race**	**Subject Reference Start Date**	**Baseline Weight (kg)**	**Baseline Height (cm)**	**Baseline BMI (kg/m2)**
012345-1560007-05006	1560007-05006	1560007	AUS	23	F	WHITE	2021-08-25	73	172	24.67550027
012345-1560007-05014	1560007-05014	1560007	FRA	21	F	WHITE	2022-05-30	42	170	14.53287197
012345-1560007-05008	1560007-05008	1560007	CHN	27	M	WHITE	2021-09-16	54	160	21.09375
012345-1560007-05004	1560007-05004	1560007	AUS	22	M	BLACK OR AFRICAN AMERICAN	2021-08-10	73	170	25.25951557
012345-1560002-05011	1560002-05011	1560002	CHN	20	M	ASIAN	2022-04-22	66.1	167	23.70110079
012345-1560002-05003	1560002-05003	1560002	CHN	36	F	ASIAN	2021-08-16	67.6	173.5	22.4567931
012345-1560006-05001	1560006-05001	1560006	CHN	19	M	ASIAN	2021-09-15	73.6	174	24.30968424
012345-1560011-05007	1560011-05007	1560011	CHN	13	M	ASIAN	2021-10-22	60	171	20.51913409
012345-1560004-05006	1560004-05006	1560004	KOR	12	M	ASIAN	2022-04-19	41.1	159	16.2572683
012345-1560004-05001	1560004-05001	1560004	KOR	13	M	ASIAN	2021-08-10	52.6	165	19.3204775

After data anonymization:

USUBJID	SUBJID	SITEID	REGIONDI	AGEDI	SEX	RACEDI	RFSTDT	WGTBLDI	BMIBLDI
Unique Subject Identifier	Subject Identifier for the Study	Study Site Identifier	De-identified Region Group	De-identified Age Band	Sex	De-identified Race Group	Subject Reference Start Date Shifted	De-identified Baseline Weight (kg)	De-identified BMI Band
012345-9990015-00016	9990015-00016	9990015	Rest of the world	[18,40)	F	WHITE	2021-07-07	[70,80)	Normal Weight
012345-9990015-00100	9990015-00100	9990015	Rest of the world	[18,40)	F	WHITE	2022-07-13	[40,50)	Underweight
012345-9990015-00177	9990015-00177	9990015	CHN	[18,40)	M	WHITE	2022-01-30	[50,60)	Normal Weight
012345-9990015-00245	9990015-00245	9990015	Rest of the world	[18,40)	M	UNKNOWN	2022-01-22	[70,80)	Overweight
012345-9990041-00122	9990041-00122	9990041	CHN	[18,40)	M	ASIAN	2022-08-19	[60,70)	Normal Weight
012345-9990041-00172	9990041-00172	9990041	CHN	[18,40)	F	ASIAN	2021-10-06	[60,70)	Normal Weight
012345-9990052-00167	9990052-00167	9990052	CHN	[18,40)	M	ASIAN	2021-05-01	[70,80)	Normal Weight
012345-9990065-00093	9990065-00093	9990065	CHN	[10,15)	M	ASIAN	2021-04-08	[50,70)	Normal Weight
012345-9990078-00114	9990078-00114	9990078	KOR	[10,15)	M	ASIAN	2021-08-01	[30,50)	Normal Weight
012345-9990078-00120	9990078-00120	9990078	KOR	[10,15)	M	ASIAN	2022-02-21	[50,70)	Normal Weight

Notes:

1. The study participant IDs were randomly generated for each study participant. New random study participant ID should begin with “999” and should have same length as original IDs.
2. Site ID is replaced by a new random site ID. New site ID should begin with "999" and have same length as original site ID.

B. The following is an example for dataset specifications:

Metadata are regenerated for anonymized data based on original metadata, adding information on anonymization rules, and dropping variables which have been dropped in the anonymization process.

Dataset Name	Variable Name	Variable Label	Origin	Derivation Type	Source	Derivation	Assigned	Define Comments	DEID_Rule
ADAE	STUDYID	Study Identifier	Predecessor		ADPSL.STUDYID				
ADAE	USUBJID	Unique Subject Identifier	De-identified		ADPSL.USUBJID				Replace the original USUBJID with a new random USUBJID.
ADAE	SUBJID	Subject Identifier for the Study	De-identified		ADSL.SUBJID				Replace the original SUBJID with a new random SUBJID.
ADAE	RACEDI	De-identified Race Group	De-identified						If the number any RACE is less than 2, RACE will be grouped into 'OTHER'.
ADAE	RANDFL	Randomized Population Flag	Predecessor		ADSL.RANDFL				
ADAE	REGIONDI	De-identified Region Group	De-identified						Create grouped COUNTRY by Standard country or area codes for statistical use (M49).
ADAE	RFICDT	Date of Informed Consent Shifted	De-identified		ADSL.RFICDT				Replace the original xxDTC with a new xxDTC transformed using the date offset method with a randomly generated number of days.
ADAE	SAFFL	Safety Population Flag	Predecessor		ADSL.SAFFL				
ADAE	ADURC	Analysis Duration (Char)	Derived	Compu-tation		Equal to "n days" (n =ADAE.ADURN) if ADURN >1; else "< 1 day" if ADURN = 1.			

Appendix: List of HIPAA Identifiers

1. Names;

2. All geographical subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code, if according to the current publicly available data from the Bureau of the Census:

(1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people, and

(2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;

3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;

4. Phone numbers;

5. Fax numbers;

6. Electronic mail addresses;

7. SSN;

8. Medical record numbers;

9. Health plan beneficiary numbers;

10. Account numbers;

11. Certificate/license numbers;

12. Vehicle identifiers and serial numbers, including license plate numbers;

13. Device identifiers and serial numbers;

14. Web Universal Resource Locators (URLs);

15. Internet Protocol (IP) address numbers;

16. Biometric identifiers, including finger and voice prints;

17. Full face photographic images and any comparable images; and

18. Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data).