Clinical Trial Data Sharing – Anonymization Standards

November 17, 2022

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 on K-ANONYMITY

  The risk is called internal risk, i.e., limited to study data. Implementation is done according to recommendations from PhUSE.

  The risk methodology is based on k-anonymity and uses the five quasi-identifiers Body Mass Index (BMI), AGE, SEX, RACE and REGION to calculate the risk. The minimum is set to k=3.

  The percentage of non k-anonymity records must be less or equal to 5% of the study population at the end of anonymization process. Then, more than 95% of the study participants cannot be distinguished with respect to these five quasi-identifiers variables BMI, AGE, SEX, RACE and REGION.

  Each release of data is such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents.

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

- 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--'.

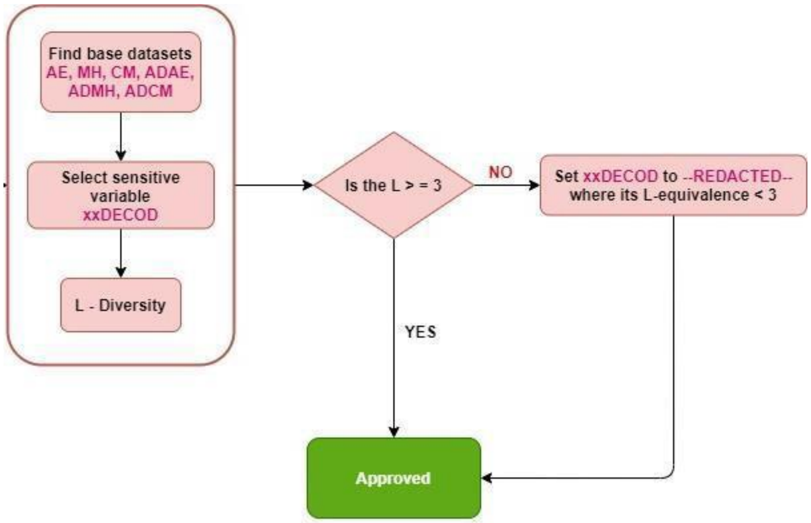

Find base datasets
AE, MH, CM, ADAE, ADMH, ADCM
Select sensitive variable
xxDECOD
L - Diversity
Is the L > = 3
NO
Set xxDECOD to --REDACTED-- where its L-equivalence < 3
YES
Approved

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 managed according to the guidance below. ● All other 16 identifiers are removed.
Study Participant Status	● Screen Failure study participants and similar study participants are dropped from datasets, as well as those for which there is no Informed Consent.
Study Participant ID Replacement	● Replace the original study participant ID with a new random study participant ID that cannot be linked back to the original study participant ID. ● Link key between actual initial study participant identifier and new random number is systematically deleted, which turns de-identification into anonymization.
Dates	● Replace all original dates by study days relative to a “baseline/ reference date” that will be provided by statisticians.
Birth date / AGE	● Date of birth is removed if it is part of the datasets. Only age is kept and converted into four classes based on quartiles. If more than 5% of the population can be distinguished on the 5 variables defined for the risk, then grouping is automatically done into two classes.
Death date	● Date of death is removed ● Relative day of death is converted to relative WEEK of death.
Free Text, Verbatim, Comments	● All free text, verbatim and comments are removed.
Reported Terms: **(Adverse Event (AE),** **Medication,** **Medical History (MH)**	● Reported AE terms by the investigator are dropped, only coded AE terms are kept, except Lowest Level Term which is dropped because too detailed. Coding hierarchy is kept starting from the Preferred Term. ● Same for MH. ● Medication: only coded transcription in international name is kept. ● Apply L- DIVERSITY to protect sensitive information on medication, MH and AE.
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 and name.
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	● For multi-country studies, a predefined model of grouping countries is applied until all combinations of GENDER * RACE * GROUPED_COUNTRY are >= 3 study participants. If a combination is still under 3, then group on RACE (see below in Demographic Information block). ● Remove results in original units. Results in standard units are kept. ● Grouping of countries does not apply in case of one country only ● Generalization on region has priority on generalization on race.
Demographic Information	● Ethnicity is removed. ● Aggregate races with few study participants under “Other” race group until all combinations of GENDER * RACE * GROUPED_COUNTRY are >= 3 study participants. ● NOT REPORTED” modality has to be kept and aggregate rule will not be applied for this group.
Weight / Height	● WEIGHT and HEIGHT are removed and relevant clinical information is given by BMI.

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.

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

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.

Data Field	Recommendation
Remove any other uniqueness of study participant record	• Aggregate fields with few study participants under a group depending on the study design
Sensitive Data (e.g., rare events, substance use)	• Check that the data do not contain specific personal data able to identify a study participant, for example some exceptionally rare AEs, or very specific substance use.

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

A more conservative approach is recommended for studies with study participant numbers below 100 and may also be considered for studies in rare diseases. Quasi-identifiers such as sex, race, BMI, weight, height, and country are removed in order to protect personal data.

Clinical trials with small duration

If the preferred approach (relative days) for dates is not used, 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:

Site ID	Investigator Name	Unique Study participant ID	Country	Race	Age (yr)	Visit Date	Weight (kg)	Height (cm)
00051	Dr. Grant	051-001	France	Caucasian	53	24JAN2011	50	170
00051	Dr. Grant	051-002	France	Caucasian	76	11FEB2011	48	140
00051	Dr. Grant	051-003	France	Black	88	03APR2010	89	182
00051	Dr. Grant	051-004	France	Caucasian	44	15AUG2011	66	178
00051	Dr. Grant	051-005	France	Caucasian	90	09SEP2011	40	155
00051	Dr. Grant	051-006	France	Black	43	21MAR2011	46	160
00051	Dr. Grant	051-007	France	Caucasian	83	25NOV2010	55	174
00052	Dr. Wilson	052-001	Spain	Asian	63	12DEC2010	87	173
00052	Dr. Wilson	052-002	Spain	Caucasian	86	07OCT2011	66	175

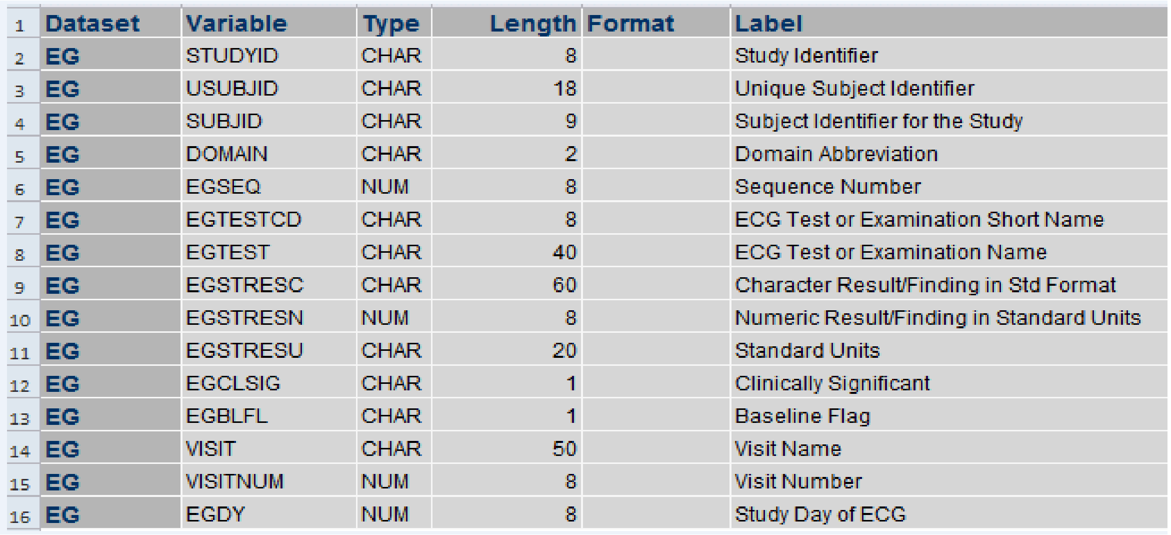
After data anonymization:

Site ID	Unique Study Participant ID	Region	Race	De-identified Age	Visit Date Relative days	Deidentified BMI (kg/m^2)
99901	999010001	Western Europe	Caucasian	[50, 70]	661	Underweight
99901	999010002	Western Europe	Caucasian	[70, 95]	679	Normal Weight
99901	999010003	Western Europe	Other	[70, 95]	365	Pre-obesity
99901	999010004	Western Europe	Caucasian	[30, 50]	864	Normal Weight
99901	999010005	Western Europe	Caucasian	[70, 95]	889	Underweight
99901	999010006	Western Europe	Other	[30, 50]	717	Underweight
99901	999010007	Western Europe	Caucasian	[70, 95]	601	Underweight
99902	999020001	Western Europe	Other	[50, 70]	618	Pre-obesity
99902	999020002	Western Europe	Caucasian	[70, 95]	917	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 length of 9.
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:



1	Dataset	Variable	Type	Length	Format	Label
2	EG	STUDYID	CHAR	8		Study Identifier
3	EG	USUBJID	CHAR	18		Unique Subject Identifier
4	EG	SUBJID	CHAR	9		Subject Identifier for the Study
5	EG	DOMAIN	CHAR	2		Domain Abbreviation
6	EG	EGSEQ	NUM	8		Sequence Number
7	EG	EGTESTCD	CHAR	8		ECG Test or Examination Short Name
8	EG	EGTEST	CHAR	40		ECG Test or Examination Name
9	EG	EGSTRESC	CHAR	60		Character Result/Finding in Std Format
10	EG	EGSTRESN	NUM	8		Numeric Result/Finding in Standard Units
11	EG	EGSTRESU	CHAR	20		Standard Units
12	EG	EGCLSIG	CHAR	1		Clinically Significant
13	EG	EGBLFL	CHAR	1		Baseline Flag
14	EG	VISIT	CHAR	50		Visit Name
15	EG	VISITNUM	NUM	8		Visit Number
16	EG	EGDY	NUM	8		Study Day of ECG

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).