Clinical Trial Data Sharing – Anonymization Standards

April 2, 2020

Table of Contents

1. CLINICAL TRIAL DATA SHARING COMMITMENT

Making clinical trial data, information and results available to researchers promises to advance science and medicine, contribute to improvements in public health and improve knowledge about, and trust in, pharmaceutical product development.

2. PURPOSE

The purpose of this document is to provide technical guidance for the de-identification of participant data from clinical trials.

This document prescribes a minimum set of steps required to protect the privacy of the participants in clinical trials by de-identifying their personal data prior to sharing 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 data de-identification steps must be taken. This trial specific approach is especially important for clinical trials in rare diseases, pediatric clinical trials and clinical trials with small sample sizes.

3. DATA AND DOCUMENTS CONSIDERED FOR DATA PREPARATION

Depending on availability, the following patient/subject level clinical trial data and related dataset documentation will be considered and prepared:

- Raw datasets or SDTM (Study Data Tabulation Model)
- Analysis datasets (ADS)
- Data set specifications including information which data dictionary was used to code some information as adverse event or medication
- Annotated case report form only, if specifically requested and available

In addition, these guidelines are used as a reference when redacting clinical trial documents, particularly in the absence of an anonymization plan, to address any participant identifiers, dates and narratives in the following clinical trial documents:

- The body of the clinical study report
- Appendices:
  - Protocol and protocol amendments
  - Sample case report form
  - Statistical Analysis Plan (SAP)/Documentation of statistical method

4. TERMS AND DEFINITIONS

The following terms are used in this document and are defined as follows:

- Personal information/data:

  Any information/data that relates to an identified or identifiable natural person and that alone or in combination with other information/data can identify an individual.

  There are two types of identifiers:

  - Direct identifiers:
    Variables that can be used alone to uniquely identify an individual;
    Examples: Names, email addresses, telephone numbers, SSN, biometrics
  - Indirect identifiers:
    Variables that can be used in combination with one another to identify an individual;
    Examples: Gender, date of birth or age, geographic locations (e.g. postal code), ethnic origin, visible minority status.

- Anonymization/de-identification:

  The process of removal or transformation of direct and indirect patient identifiers including the destruction of the key link to rendering data into a form which does not identify individuals and where identification is not likely to take place; transformation techniques include but are not limited to suppression, generalizing, or replacing with random values.
  The terms anonymization and de-identification can be used interchangeably. However, using the term anonymization presumes the destruction of the key link.

- Pseudonymization:

  The process of replacing one attribute (typically a unique attribute) by another. The natural person is still likely to be identified indirectly. Pseudonymization reduces the possibility to link a dataset with the original identity of a data subject.

- Redaction:

  The overall process of modifying a clinical trial document from its original state to protect personal information as well as proprietary and confidential information.
  In addition, the term redaction is used to describe a specific technique to modify content from its original form.

- Commercially Confidential Information:

  Any information such as a specific assay or data collection method contained in the clinical documents that is not in the public domain or publicly available and disclosure of which may undermine the legitimate economic interest of the company must be redacted.

5. MANAGEMENT OF THE DE-IDENTIFICATION PROCESS

5.1 STANDARD DATA FIELDS CONSIDERED FOR DE-IDENTIFICATION

The following types of identifiers are examples, as identified by HIPAA/Safe Harbor, that are considered for removal or transformation when anonymizing/de-identifying patient level clinical trial data and supporting trial documents to prevent the risk of association of a trial participant to his/her data. The list includes, but is not limited to:

1) Names and initials
2) All elements of dates (except year) which can be directly associated with a specific individual (birthdate, date of death, adverse event date, admission date, discharge date, etc.)
3) Kit numbers (diagnostic kits) and device numbers (devices used in the trials)
4) Geographic information such as place of work, trial site location, addresses, zip codes, etc.
5) Telephone numbers
6) Email addresses
7) Fax numbers
8) Account numbers
9) Social security numbers
10) Health plan beneficiary numbers
11) Medical record numbers
12) Vehicle identifier numbers and serial numbers including license plate numbers
13) Certificate / license numbers (marriage licenses, etc.)
14) Biometric identifiers including such as MRI, hand voice prints, etc.
15) Full face photographic images or comparable images
16) Web Universal Resource Locators (URLs)
17) Internet Protocol (IP) addresses
18) Any other unique identifying number, code or characteristic.

All of these 18 identifiers except for elements of dates and geographic information must be considered to be removed from the data sets.
The below table provides details of transforming information related to dates and geographic information and identifies several attributes which should be addressed when de-identifying data related to clinical trial participants. The information is not exhaustive. There are attributes which may be product, study phase and / or disease-specific and should be addressed accordingly.

Data Field	By Default De-Identification Approach
HIPAA 18 Identifiers	● Dates and geographic information are managed according to the guidance below. ● All other 16 identifiers are removed.
Subject ID Replacement	● Replace the original subject ID with a new random subject ID that cannot be linked back to the original subject ID.
Dates	● Replace all original dates by study days relative to a “baseline/ reference date” that will be provided by statisticians. ● Replace all original dates related to individual subjects with new dates offset by delta. Delta is defined for each patient as first date – trial initiation date (patient specific FPFV – trial FPFV). ● Other publicly available dates such as motor vehicle accidents, country- and religion-specific holidays must be transformed in the same way as the other dates; if transforming is not possible, then these dates must be removed.
Birth date / AGE	● Date of birth is removed if it is part of the datasets. Only age is kept. ● If some outliers are identified on the AGE, then those outliers are set to blank, and replaced by a group containing at least 3 patients/subjects. If there is no outlier, then no grouping is applied. ● In addition to the check on outliers, a threshold value at 84 years old is applied. This means that whatever outliers, all ages above 84 are set to blank and replaced by a group >=84.
Death date	● Relative day of death is converted to relative WEEK of death, in order to protect privacy. ● In addition, date of death is removed.
Free Text, Verbatim, Comments	● All free text, verbatim and comments are removed.
Reported Terms (Adverse Event,	● Reported terms on AE, concomitant medications and medical history are removed; only coded terms are kept.

Data Field	By Default De-Identification Approach
Medication, **Medical History)**	
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 patients, e.g. pacemaker, to minimize risk of re-identification.
Geographic Information	● For multi-country studies, a predefined model of grouping countries is applied until all combinations of GENDER * RACE * GROUPED_COUNTRY are >= 3 patients/subjects. 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
Investigator ID & Name	● Remove investigator identifier and name.
Site ID	● Site ID is replaced in a similar mode as for Subject 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 de-identified study datasets.
Demographic Information	● Ethnicity is removed. ● Aggregate races with few patients under “Other” race group until all combinations of GENDER * RACE * GROUPED_COUNTRY are >= 3 patients/subjects. ● “NOT REPORTED” modality has to be kept and aggregate rule will not be applied for this group.
Genetic Data	● Remove all genetic data.

Data Field	By Default De-Identification Approach
Weight / Height	● If some outliers are identified on weight or height by gender, then those outliers are set to blank, and replaced by a group containing at least 3 patients/subjects, by gender. If there is no outlier, then no grouping is applied. ● Remove height, weight, BMI and age completely, no transforming, for studies with less than 100 patients.

Screening failure subjects and similar subjects are excluded from sharing, as well as those for which there is no Informed Consent.

Please also note: The above rules should be used as a reference when redacting clinical trial documents and applied to any participant identifiers, dates and narratives if appropriate.

5.2 NON-STANDARD APPROACH CONSIDERED FOR DE-IDENTIFICATION

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 Patient record	● Aggregate fields with few patients 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 patient, for example some exceptionally rare AEs, or very specific substance use.

Clinical trials with small sample size and all clinical trials in rare diseases:
A more conservative approach by removing identifier and quasi identifier is recommended 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.

5.3 DE-IDENTIFICATION EXAMPLES

A. The following example shows a dataset before and after the data de-identification process applying the minimum set of data de-identification steps.

Before data de-identification:

Site ID	Investigator Name	Unique Subject 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 de-identification:

Site ID	Unique Subject ID	Region	Race	Age (yr)	Deiden-tified Age	Visit Date Relative days	Weight (kg)	Deidenti-fied Weight (kg)	Height (cm)	Deidentified Height (cm)
99901	999010 001	Western Europe	Cau-casian	53	53	661	50	50	170	170
99901	999010 002	Western Europe	Cau-casian	76	76	679	48	48		<156
99901	999010 003	Western Europe	Other		>84	365		>86		>179
99901	999010 004	Western Europe	Cau-casian		<45	864	66	66	178	178
99901	999010 005	Western Europe	Cau-casian		>84	889		<47		<156
99901	999010 006	Western Europe	Other		<45	717		<47	160	160
99901	999010 007	Western Europe	Cau-casian	83	83	601	55	55	174	174
99902	999020 001	Western Europe	Other	63	63	618		>86	173	173
99902	999020 002	Western Europe	Cau-casian		>84	917	66	66	175	175

Notes:

1. The subject IDs were randomly generated for each subject. New random subject 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.
3. All standard dictionary coded terms will be retained and reported terms are removed. Yet coded terms may be removed if the requestor is not allowed to use a specific dictionary; in this case we keep the reported terms provided it does not contain any personal data.

B. The following is an example for dataset specifications.

	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

End of Document