Novartis Global Clinical Data Anonymization Standards (version 4)

Contents

1 Introduction

Patient-level data collected in Novartis clinical trials will be anonymized according to the standards set forth in this document. These standards will ensure compliance with current privacy laws and regulatory guidance while allowing data to be shared with researchers. There are a number of data elements enumerated in the "Privacy Rule" under the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and other guidance from European General Data Protection Regulation which can be used to identify individuals. The process of anonymizing can be thought of as permanently removing the ability to use any of these elements to identify individual participants. Direct and indirect identifiers are transformed thereby making it unlikely to allow any individual to be identified by combining data. To that end, a risk-based approach is implemented to determine the required transformations on indirect identifiers and to ensure that risk of re-identification of study participants is very low. Adherence to the framework of these standards will minimize the risks of encroaching on the privacy and confidentiality of research participants.

2 General Approach

Upon approved requests, the following data and accompanying trial documentation will be shared with qualified external researchers when available.

2.1. Original documentation and amendments that articulate statistical methodology
2.2. CSR (Redacted) appendices
2.3. Annotated CRF
2.4. Dataset specifications
2.5. Anonymized raw study datasets – collected data from each patient in the study
2.6. Anonymized analysis-ready datasets – data used for analysis
This document describes the principles for generating the anonymized datasets. There will be no way to undo and recreate the original data once it is anonymized.

We advise researchers to send queries they may have on the anonymized data as early as possible as investigations are faster conducted on recently anonymized data.

3 Anonymizing Personally Identifiable Information (PII)

3.1 PII

The way the variables are handled during risk measurement and anonymization process will depend on how they are classified. A distinction is made among three types of variables.

- Directly identifying variables:
  One or more direct identifiers can be used to uniquely identify an individual, either by themselves or in combination with other readily available information, for example subject ID.
- Indirectly identifying variables (quasi-identifiers):
  The quasi-identifiers are the background knowledge variables about individuals in the disclosed data set that an adversary can use, individually or in combination, to probabilistically re-identify a trial participant. If an adversary does not have background knowledge of a variable, then it cannot be a quasi-identifier. The manner in which an adversary can obtain such background knowledge will determine which attacks on a data set are plausible. For example, the background knowledge may be available because the adversary knows a particular target individual in the disclosed data set, an individual in the data set has a visible characteristic that is also described in the data set, or the background knowledge exists in a public or semi-public registry. Examples of quasi-identifiers include sex, date of birth or age, locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, aboriginal identity, total years of schooling, marital status, criminal history, total income, visible minority status, event dates (such as admission, discharge, procedure, death, specimen collection, visit/encounter), codes (such as diagnosis codes, procedure codes, and adverse event codes), and country of birth.

- Other variables:
  These are the variables that are not useful for determining an individual's identity. They may be clinically relevant or not. Examples of clinical variables are numerical laboratory test results and drug dosage information.

This will be used as a framework for defining the Novartis anonymization standards, discussed in the following sections.

3.2 Identifiers

Change the real value to a de-identified value in a consistent manner so that the value in one instance of the variable is consistent with the value in the same variable across other datasets. This does not limit but includes PK datasets and central lab data. Extension studies use the same new identifiers as were used in the initial study to preserve the links between studies. This also applies to long-term follow-up studies where separate reports are published.

3.2.1. The investigator number is set to blank for each investigator. The investigator name is set to blank in the dataset.
3.2.2. Each participant is given a new subject identifier.
3.2.3. Center numbers and randomization numbers are masked.

3.3 Free Text Verbatim Terms

Information in a descriptive free text verbatim term may compromise a participant's anonymity. Therefore, these fields are not included in risk of re-identification determination and are set to blank or investigated for PII (if possible).

3.4 Date of Birth

Information relating to a research participant's date of birth and age may compromise anonymity. Date of birth is suppressed (or generalized if age does not exist in the data) and age is generalized based on the outcome of risk of re-identification determination.

3.5 Other Dates

Specific dates directly related to a research participant may compromise a research participant's anonymity.

Therefore, dates are shifted according to the scheme defined in the PhUSE CDISC SDTM anonymization standard. This scheme determines an offset for each subject based on a difference between a date in the trial available for all patients (for example screening date) and an anchor date which is typically study initiation date.

For each subject, the offset is applied to all subject's dates. All original dates are replaced

with the new dummy dates so that the relative times between dates are retained.

Example: If the original reference date was 01APR2020 and the date of death was 01MAY2020, and offset is 91 days. Dummy dates are then calculated using this offset of 91 days.

	Original Date	New Date	
Reference Date	01APR2020	01JAN2020	Apply offset = 91 days
Date of Death	01MAY2020	31JAN2020	Apply offset = 91 days
Relative Time of Death	30 days	30 days	

3.6 Other PII

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

3.6.1. Information from variable names e.g. lab names may contain location information
3.6.2. Investigator comments may be used to identify subjects
3.6.3. Genetic data will not be shared at all
3.6.4. Also excluded will be case narratives, documentation for adjudication and images (e.g. x-rays, MRI scans)

4 Remnants

After anonymization, there is no information available that will allow us to recreate the original datasets from the anonymized data. This includes but is not limited to the following:

4.1 Any transactional copies of anonymized datasets
4.2 De-identification tables (links from original variable to new anonymized variable)
4.3 QC output datasets
4.4 Any Log files
4.5 The seed utilized for random number generation

The anonymized datasets are stored separately from the original datasets in the Novartis systems.

5 Example

Study data example on top and anonymized data in the 2nd set of rows.

Center ID	Investigator ID	Investigator name	Subject number	Date of birth	Age (yrs)	AE start date	AE end date	Verbatim term	Preferred term
T1230	279T344	Dr Smith	2002	08Aug1954	57	29DEC2010	27JAN2011	HEADACHE	Headache
T1230	279T344	Dr Smith	2002	08Aug1954	57	10JAN2011	06APR2011	BRONCHITIS	Bronchitis
T1230	279T344	Dr Smith	2004	09Aug1919	92	25MAR2011	12AUG2011	COLD	Nasopharyngitis
T1230	279T344	Dr Smith	2004	09Aug1919	92	28MAR2011	31MAR2011	FLU	Influenza
T1230	279T344	Dr Smith	2004	09Aug1919	92	01MAR2011	15MAY2011	PAIN	Pain
G5670	348G224	Dr Jones	2010	09Aug1947	64	14OCT2010	20OCT2011	ACHE NOS	Pain
G5670	348G224	Dr Jones	2010	09Aug1947	64	24MAY2011		BRONCHIAL INFECTION	Bronchitis
G5670	348G224	Dr Jones	2010	09Aug1947	64	01MAR2011	15MAR2011	CHRONIC PAIN	Pain
Replace	Replace	Set to blank	Replace	Drop	Aggregate ages >=90	Replace	Replace	Set to blank	Keep
Center ID	Investigator ID	Investigator name	Subject number		Age (yrs)	AE start date	AE end date	Verbatim term	Preferred term
Xnn10	nnnXn10		1111		57	19AUG2010	17SEP2010		Headache
Xnn10	nnnXn10		1111		57	06JUL2010	20SEP2010		Bronchitis
Xnn10	nnnXn10		1113		90 or Older	05SEP2010	23JAN2011		Nasopharyngitis
Xnn10	nnnXn10		1113		90 or Older	06SEP2010	09SEP2010		Influenza
Xnn10	nnnXn10		1113		90 or Older	29JUN2011	12SEP2011		Pain
Xnn11	nnnXn11		1101		64	16JUL2011	12SEP2011		Pain
Xnn11	nnnXn11		1101		64	04NOV2010			Bronchitis
Xnn11	nnnXn11		1101		64	01JUL2010	15JUL2010		Pain

6 Reference List

Guidance on De-identification of Protected Health Information – US
http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule
http://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf

European Union General Data Protection Regulation
http://data.europa.eu/eli/reg/2016/679/oj