Novartis Global Data Anonymization Standards

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 removed thereby making it unlikely to allow any individual to be identified by combining data. 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. This document will focus on the last two points below, 2.6 and 2.7.

2.1. Original protocol and any amendments
2.2. Original documentation and amendments that articulate statistical methodology
2.3. CSR (Redacted) appendices
2.4. Annotated CRF
2.5. Dataset specifications

2.6. Anonymized raw study datasets – collected data from each patient in the study
2.7. Anonymized analysis-ready datasets – data used for analysis

The raw and analysis ready datasets will be anonymized where all personally identifiable information (PII) will be removed or replaced. Subject identifiers will be recoded. Free text will be removed. Date of birth will be dropped; age will be categorized. Dates will be offset to a point into the future. There will be no way to undo and recreate the original data once it is anonymized per section 4.

3 Removing Personally Identifiable Information (PII)

3.1 PII

There are 18 identifiers to be removed from the datasets (and related documentation) as described in (HIPAA) CFR – Title 45: Public Welfare, Subtitle A §164.514. The identifiers to be removed are:

3.1.1. Names
3.1.2. All geographic subdivisions smaller than a state including:
Street address, City, County, Precinct, Zip Code and Geocodes
Except for the initial 3 digits of a zip code if:
3.1.2.1. The geographic area formed by combining all zip codes with the same 3 initial digits contains more than 20,000 or fewer people and
3.1.2.2. The initial 3 digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
3.1.3. 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
3.1.4. Telephone numbers
3.1.5. Fax numbers
3.1.6. Electronic mail addresses
3.1.7. Social security numbers
3.1.8. Medical record numbers
3.1.9. Health plan beneficiary numbers
3.1.10. Account numbers
3.1.11. Certificate/license numbers
3.1.12. Vehicle identifiers and serial numbers, including license plate numbers
3.1.13. Device identifiers and serial numbers
3.1.14. Web Universal Resource Locators (URLs)
3.1.15. Internet Protocol (IP) address numbers
3.1.16. Biometric identifiers, including finger and voice prints
3.1.17. Full face photographic images and any comparable images
3.1.18. Any other uniquely identifying number, characteristic, or code

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 re-coded or set to blank for each investigator. The investigator name is set to blank or dropped from the dataset.
3.2.2. Each participant is given a new subject identifier.
3.2.3. Each center is given a new identifier. Trials containing one or more center with <10 patients will need to be dealt with on a trial by trial basis. Aggregation of centers can be considered or possibly dropping center.

3.3 Free Text Verbatim Terms

Information in a descriptive free text verbatim term may compromise a participant's anonymity.

3.3.1. Free text verbatim terms are not included and are set to blank or dropped from the following datasets:
3.3.1.1 Adverse Events
3.3.1.2 Medications
3.3.1.3 Medical History
3.3.1.4 Other specific verbatim free text
3.3.2. All standard dictionary coded terms will be retained.

3.4 Date of Birth

Information relating to a research participant's date of birth and identification of specific ages above 89 may compromise anonymity.

Date of birth is dropped and ages above 89 are aggregated into a single category of "90 or older". (See section 5 example)

3.5 Other Dates

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

3.5.1. A random offset per study, is generated and applied to all dates. All original dates are replaced with the new dummy dates so that the relative times between dates are retained.
3.5.2. This date offset will be generated at the study level pushing dates into the future.
3.5.3. Do not retain any seasonal information.

Example: If the original reference date was 01APR2008 and the date of death was 01MAY2008, a random offset is generated (in this case 91 days). Dummy dates are then calculated using this offset of 91 days.

	Original Date	New Date	
Reference Date	01APR2008	01JUL2008	Apply offset = 91 days
Date of Death	01MAY2008	31JUL2008	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 a subject
3.6.3. Genetic data will be not be shared at all
3.6.4. Exploratory Biomarker data outside the primary and key secondary endpoints and laboratory data
3.6.5. Also excluded will be case narratives, documentation for adjudication and imaging data (e.g. x-rays, MRI scans)

4 Remnants

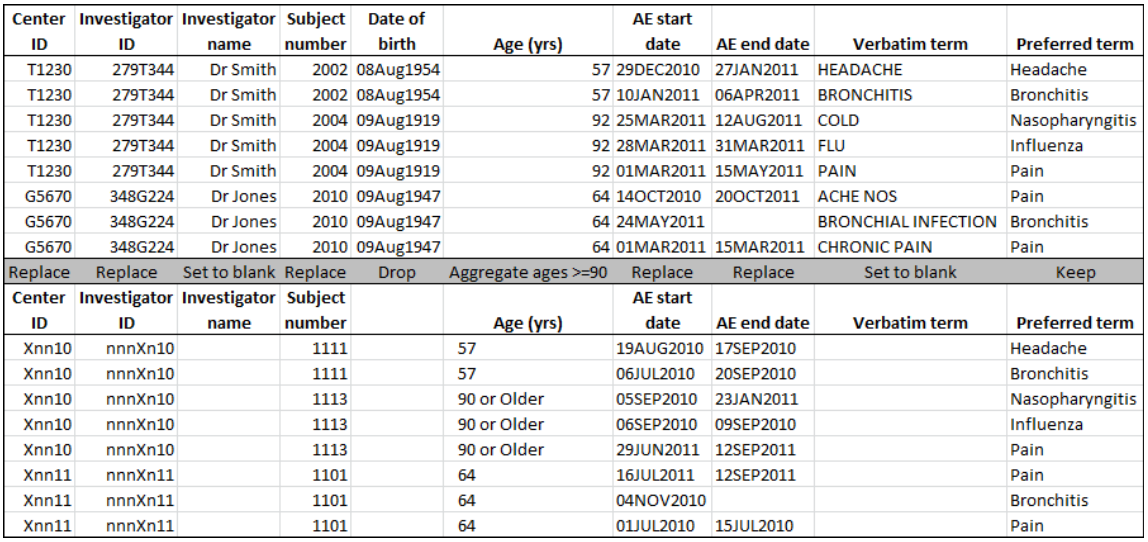
After anonymization, there is no information available that will allow us to recreate the original datasets from the anonymized data. This includes but is not limited to the following:

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 or LST 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://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:008:0001:0022:en:PDF