Anonymisation of Clinical Trial Datasets

1. Introduction

Providing access to data in ways that allows further research while maintaining the privacy and confidentiality of research participants is critical. There are also privacy laws and regulatory guidance which need to be followed (for example guidance from European data protection regulators and Code of Federal Regulations - Title 45: Public Welfare, Subtitle A §164.514). Publications in this area which provide guidance[1],[2].

This document describes the approach taken by Eisai to prepare data for sharing with other researchers in a way that:

- Minimises risks to the privacy and confidentiality of research participants.
- Ensures compliance with data privacy legal requirements.

Other study sponsors achieve these objectives using other approaches (see the Study sponsors section of ClinicalStudyDataRequest.com).

2. General Approach

Access is provided to anonymised data. Anonymisation involves:

a. **Removing personally identifiable information (PII) from the dataset.** This includes recoding identifiers (by replacing the original code number with a new code number), removing free text verbatim terms, Replacing date of birth with age and replacing all dates relating to individual subjects with dummy dates.

b. **Destroying the link (code key) between the dataset that is provided and the original dataset.** Some Data Protection Authorities in Europe suggest that the data can only be considered anonymised if personal information is removed (or redacted) and the subject code number cannot be linked to a research participant. Therefore, research participants' identification code numbers are anonymised by destroying the code key that was used to generate the new code number from the original (i.e. destroying the link between the two code numbers).

[1] Hrynaszkiewicz I, Norton ML, *et al.* Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ* 2010; **340**: c181.

[2] De-identification of Clinical Trials Data Demystified. *Jack Shostak, Duke Clinical Research Institute (DCRI), Durham, NC http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf*

3. Removing personally identifiable information (PII) from the dataset

The 18 identifiers (as defined by HIPAA –see Code of Federal Regulations - Title 45: Public Welfare, Subtitle A §164.514) are removed from the datasets (and related documentation). In addition any other PII that may be present is removed.

This involves removing:

- any names and initials,
- kit numbers and device numbers
- geographic information such as place of work.

In addition the following steps are undertaken:

- Recoding identifiers (or code numbers).
- Removing free text verbatim terms.
- Removing date of birth and using age at randomisation. Ages above 89 which are aggregated into a single category of "90 or older". (This is a specific HIPAA requirement).
- Replacing all original dates relating to individual subjects with offsets which are then applied to create 'dummy dates'. (see below)
- Reviewing and removing other PII

These steps are described in further detail below.

3.1 Recoding Identifiers (or code numbers)

The following identifiers (code numbers) are re-coded and the code key that was used to generate the new code number from the original code number is destroyed (as described in section 5):

    - The investigator identifier (or code number) is re-coded or set to blank for each investigator. The investigator name is set to “blank” or dropped from the dataset (see Appendix 1).
    - A new subject identifier (or code number) for each research participant.
    - Re-code or blank the centre identification number.
    - Aggregate patients from centres with less than 12 patients into a single centre.

- The same new identifiers (or code numbers) are used across all datasets applicable to a single study e.g. raw dataset, analysis-ready dataset. This includes (where applicable) PK datasets, genetic datasets etc.

- Extension studies use the same new identifiers (or code numbers) as used for the initial study to enable individual subject data to remain linked. This also applies to long term follow-up studies where separate reports are published. This is achieved by repeating the data anonymisation process for the initial study data at the same time as the extension/follow up data.

3.2 Removing Free Text Verbatim Terms

Information in a descriptive free text verbatim term may compromise a subject’s anonymity.

- Free text verbatim terms are set to “blank” or dropped from the dataset including:
    - Adverse Events
    - Medications
    - Other e.g. Medical History
    - Other specific verbatim free text

  Certain free text fields may be retained if they do not contain PII and removal of these fields may impact the scientific value of the dataset (e.g. medical history that has not been coded).

- All dictionary coded terms with decode and/or terms that use a pre-specified list are retained.

3.3 Replacing 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 blanked, age at randomisation is kept with the exception of ages above 89 which are aggregated into a single category of “90 or older”

3.4 Replacing all Original Dates relating to a Research Participant

Eisai uses the methods described below.

3.4.1 Dummy Date Method

Specific dates (other than year) directly related to a research participant may compromise a research participant’s anonymity.

All dates are replaced. An offset is created for each research participant and applied to all dates for that research participant, moving the subject’s first on-study date (consent date) to the study start date (first consent date). All original dates are replaced with the new dummy dates so that the relative times for each research participant are retained.

Example: If the original reference (consent) date was 01JUL2018 and the date of death was 31JUL2018, and the study start date was 01APR2018: an offset of 91 days is generated. Dummy dates are then calculated using this offset of 91 days.

	New Date	Original Date	
Reference date	01Apr2018	01Jul2018	Apply offset = 91 days
Date of Death	01May2018	31Jul2018	Apply offset=91 days
Relative Time of death	30 days	30 days	

3.5 Reviewing and Removing Other PII

- Other data elements that contain PII are removed. For example:
    - Information from variable names e.g. lab names may contain location information
    - Investigator comments may be used to identify a subject
    - Genetic data that would enable a direct trace back to an individual subject

Appendix 1: Illustrates non-real examples of how these steps are applied.

4. Review and Quality Control

A final review of the HIPAA 18 identifiers is made to determine if further removal is required. Quality Control checks and documentation (QC record) is conducted for the processing of the data and supportive metadata documentation.

5. Destroying the link (key code) between the dataset that is provided and the original dataset

Research participants' identification code numbers are anonymised by replacing the original code number with a new code number (as described in 3.1) and destroying the code key that was used to generate the new code number from the original (i.e. destroying the link between the two code numbers).

The anonymised datasets are stored in a separate secure location to the original coded datasets.

Appendix 1: A non-real example illustrating removal of personally identifiable information using the dummy date method

Centre ID	Investigator ID (INVID)	Investigator name (INVNAME)	Subject ID (SUBID)	Unique subject ID (USUBID)	Age (yrs)
00123	279344	Dr Smith	5	TJF4392.005	57
00123	279344	Dr Smith	2	TJF4392.002	72
00123	279344	Dr Smith	1	TJF4392.001	91
00123	279344	Dr Smith	66	TJF4392.066	89
00123	279344	Dr Smith	8	TJF4392.008	94
05678	333721	Dr Jones	19	TJF4392.019	85
05678	333721	Dr Jones	4	TJF4392.004	53
05678	333721	Dr Jones	23	TJF4392.002	76

AE start date	AE end date	Verbatim term
29DEC2010	27JAN2011	Headache
10JAN2011	06APR2011	Nausea
25MAR2011	12AUG2011	Cold
28MAR2011	31MAR2011	Cold
01MAR2011	15MAY2011	Flu
14OCT2010	20OCT2011	Cold
24MAY2011	.	Headache
01MAR2011	15MAR2011	Pain

New INVID

Remove INVNAME

New SUBID

New USUBID

Remove ages above 89

Create age category

Add dummy dates

Add dummy dates

Remove

Centre ID	Investigator ID (INVID)	Investigator name	Subject ID (SUBID)	Unique subject ID (USUSID)	Age (yrs)	Age Category	AE start date	AE end date	Verbatim term
00123	227		8754	TJF4392.8754	57	<=89	19AUG2010	17SEP2010	
00123	227		5681	TJF4392.5681	72	<=89	06JUL2010	30SEP2010	
00123	227		1475	TJF4392.1475	.	>89	05SEP2010	23JAN2011	
00123	227		6589	TJF4392.6589	89	<=89	06SEP2010	09SEP2010	
00123	227		3562	TJF4392.3562	.	>89	29JUN2011	12SEP2011	
05678	208		1457	TJF4392.1457	85	<=89	16JUL2011	12SEP2011	
05678	208		2214	TJF4392.2214	53	<=89	04NOV2010	.	
05678	208		2236	TJF4392.2236	76	<=89	01JUL2010	15JUL2010	