Shionogi Data Anonymization Standards

1. Introduction

Providing access to data in ways that allows further research while maintaining the privacy and confidentiality of research participants is important for everyone involved in Shionogi trials. There are several privacy laws and regulatory guidance documents 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 Shionogi approach to the preparation of data for sharing with other researchers in a way that:

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

2. General Approach

Upon approval of the research proposal by the Independent Review Panel (IRP), the requested data and relevant study documents are shared with the research team. These datasets and documents may include.

- Raw study datasets or SDTM datasets
- Analysis-ready datasets
- Annotated Case Report Form (aCRF)
- Dataset specifications
- Protocol with any amendments
- Statistical Analysis Plan (SAP)
- Redacted clinical study report

Raw or SDTM datasets and analysis-ready datasets are anonymized by removing, replacing, sub-sampling [or shuffling] all Personally Identifiable Information (PII). Subject identifiers are recoded consistently across all datasets, to break any links with original study data or documentation, whilst ensuring all data of one subject remains linked together.

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

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:

(A) Names

(B) All geographic 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 or fewer 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.

(C) 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

(D) Telephone numbers

(E) Fax numbers

(F) Electronic mail addresses

(G) Social security numbers

(H) Medical record numbers

(I) Health plan beneficiary numbers

(J) Account numbers

(K) Certificate/license numbers

(L) Vehicle identifiers and serial numbers, including license plate numbers

(M) Device identifiers and serial numbers

(N) Web Universal Resource Locators (URLs)

(O) Internet Protocol (IP) address numbers

(P) Biometric identifiers, including finger and voice prints

(Q) Full face photographic images and any comparable images

(R) Any other uniquely identifying number, characteristic, or code

In addition, Shionogi will apply the following steps as described below.

3.1 Recoding Identifiers (or code numbers)

The following identifiers 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 (or site) identifier (or code number) is re-coded for each investigator (site). The investigator (site) name is set to “blank” or dropped from the dataset (see Appendix 1 & 2).

- A new subject identifier (or code number) for each research participant, which is consistently applied across all datasets in the study.
- The same new identifiers (or code number) are used across all datasets applicable to a single study e.g. raw dataset, analysis-ready dataset. This includes (where applicable) pharmacokinetic datasets, genetic datasets etc.
- Extension studies use the same new identifiers (or code number) 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 anonymization 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
  - medical history
  - other specific verbatim free text

  Certain free text fields may be retained if they do not contain PII as 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 verbatim 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 set to blank and the age at reference start date 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

Use one of two methods as 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: A random offset is generated for each research participant and applied to all dates for that research participant. 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 date was 01APR2008 and the date of death was 01MAY2008, a random offset is generated (in this case 91 days). Dummy dates are than 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	

3.4.2 Study Day Method:

All dates are set to blank. If not available in the dataset already, the study day is calculated for each observation with days relative to a reference date. If clearly defined in the dataset model or specifications, the reference date is used as defined therein. If not, in order of priority the reference date is defined as the date of first study treatment, date of randomization or date of consent. For example if a patient is randomized, but does not take the study treatment (i.e. the date of first treatment is missing), the date of randomization is used as the reference date to calculate the study day for any assessments recorded.

Example If the original reference date was 01JAN2008 and the date of death was 01MAY2008, the date of death would be 122 expressed as study days.

	Original Date	Reference Date	Study Day
Date of Death	01May2008	01Jan2008	122

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 that may be used to identify a subject
- Genetic data that may 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 assigned DI (de-identification) rules is made to determine if further removal is required. Quality Control (QC) 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

Some data protection authorities in Europe suggest that the data can only be considered anonymized 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 anonymized 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).

Research participants' identification code numbers are anonymized 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 following specific items are discarded:

- Any transactional copies of anonymized datasets
- De-identification tables (links from original variable to new anonymized variable)
- QC output datasets and review files
- Any [SAS] log or output (e.g. lst) files that contain PII
- The seed utilised for random number generation

The anonymized datasets are stored in a separate secure location from the original datasets.

References:

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

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)	AE start date	AE end date	Verbatim term
00123	279344	Dr Smith	5	TJF4392.005	57	29DEC2010	27JAN2011	Headache
00123	279344	Dr Smith	2	TJF4392.002	72	10JAN2011	06APR2011	Nausea
00123	279344	Dr Smith	1	TJF4392.001	91	25MAR2011	12AUG2011	Cold
00123	279344	Dr Smith	66	TJF4392.066	89	28MAR2011	31MAR2011	Cold
00123	279344	Dr Smith	8	TJF4392.008	94	01MAR2011	15MAY2011	Flu
05678	333721	Dr Jones	19	TJF4392.019	85	14OCT2010	20OCT2011	Cold
05678	333721	Dr Jones	4	TJF4392.004	53	24MAY2011	.	Headache
05678	333721	Dr Jones	23	TJF4392.002	76	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	

Appendix 2: A non-real example illustrating removal of personally identifiable information using the study day method and aggregation of small centres

Centre ID	Investigator ID (INVID)	Investigator name (INVNAME)	Subject ID (SUBID)	Unique subject ID (USUBID)	Age (yrs)	AE start date	AE end date	Verbatim term
00123	279344	Dr Smith	5	TJF4392.005	57	29DEC2010	27JAN2011	Headache
00123	279344	Dr Smith	2	TJF4392.002	72	10JAN2011	06APR2011	Nausea
00123	279344	Dr Smith	1	TJF4392.001	91	25MAR2011	12AUG2011	Cold
00123	279344	Dr Smith	66	TJF4392.066	89	28MAR2011	31MAR2011	Cold
00123	279344	Dr Smith	8	TJF4392.008	94	01MAR2011	15MAY2011	Flu
05678	333721	Dr Jones	19	TJF4392.019	85	14OCT2010	20OCT2011	Cold
05678	333721	Dr Jones	4	TJF4392.004	53	24MAY2011	.	Headache
05678	333721	Dr Jones	23	TJF4392.002	76	01MAR2011	15MAR2011	Pain

⇩ New centre ID. Merged as < 10 patients ⇩ Remove INVID ⇩ Drop INVNAME from dataset ⇩ New SUBID ⇩ New USUBID ⇩ Remove ages above 89 ⇩ Create age category ⇩ Calculate study day ⇩ Calculate study day ⇩ Remove

Centre ID	Investigator ID (INVID)
22265	
22265	
22265	
22265	
22265	
22265	
22265	
22265	

Subject ID (SUBID)	Unique subject ID (USUBID)	Age (yrs)	Age Category	AE start date	AE end date	Verbatim term
8754	TJF4392.8754	57	<=89	20	49	
5681	TJF4392.5681	72	<=89	15	101	
1475	TJF4392.1475	.	>89	322	462	
6589	TJF4392.6589	89	<=89	17	20	
3562	TJF4392.3562	.	>89	23	98	
1457	TJF4392.1457	85	<=89	2	373	
2214	TJF4392.2214	53	<=89	4	.	
2236	TJF4392.2236	76	<=89	15	29	