Anonymization 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). There are also a number of publications in this area which provide guidance[1,2].

UCB anonymizes data according to a quantitative approach. This document describes the approach taken by UCB to prepare 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 requirements

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

2. General Approach

Access is provided to anonymized data. The UCB anonymization process involves:

a. **De-identifying the data by removing personally identifiable information (PII).** This includes recoding identifiers (by replacing the original code number with a new code number), removing or redacting free text verbatim terms and comments, removing date of birth and replacing age with age band, and replacing all dates relating to individual research participants 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 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 (ie, 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 or redacting:

- Any names and initials
- (or recoding) batch, kit, and device numbers
- Geographic information such as place of work
- Socioeconomic data such as occupation, income or education, household and family composition, multiple pregnancies, where these data are considered to jeopardize the privacy of the research participants.

In addition, the following steps are undertaken:

- Recoding identifiers (or code numbers)
- Removing or redacting free text verbatim terms
- Date of birth is set to blank. Age band is provided and is study-specific, looking at the frequency counts of age by treatment and grouping into bands using k-anonymity.
- Date of death (if collected) is set to blank.
- Replacing all original dates relating to individual research participants with randomly generated 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):

- A new research participant identifier (or code number) for each research participant.
- A new center identifier (or code number) for each center.
- The investigator identifier (or code number) is re-coded for each investigator. The

investigator name is set to “blank” (see Appendix 1).
  - A new third party vendor identifier (or code number) for each third party vendor, eg, laboratory or radiology unit.
  - The country identifier will be set to blank. Locational information will only be provided at a regional level for trials that used region as stratification variables.

- 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 research participant 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.

The resultant datasets are sorted by the newly created research participant identifiers in order to change the order of records in the de-identified datasets compared to the original datasets.

3.2 Removing Free Text Verbatim Terms

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

- Free text verbatim terms are set to “blank” including (but not limited to):
  - Adverse Events
  - Medications
  - Other e.g. Medical History
  - Other specific verbatim free text
  - Comment fields

- Certain free text fields may be retained if they do not contain PII and where removal of these fields may impact the scientific value of the dataset (eg, medical history that has not been coded, or indication for medications).
- All dictionary coded terms with decode and/or verbatim terms that use a pre-specified list are retained.

3.3 Replacing Date of Birth

Date of birth is set to blank. Age band is provided and is study-specific, looking at the frequency counts of age by treatment and grouping into bands using k-anonymity.

3.4 Replacing all Original Dates relating to a Research Participant

Dummy Date Method

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

All dates (and date parts of 'datetimes') are offset by a constant number of days, an "offset", for each research participant. If the original date variable contains partial date values, these are imputed to a full date and offset in the same way. This ensures the relative distances between dates for each research participant are retained.

Example: If the original reference date was 01APR2008 and the onset date of an adverse event was 01MAY2008, a random offset is generated (in this case 91 days). Dummy dates are then calculated for this research participant's data using this offset of 91 days.

	Original Date	New Date	
Reference date	01APR2008	01JUL2008	Apply offset = 91 days
Adverse Event Date	01MAY2008	31JUL2008	Apply offset = 91 days
Relative Time of adverse event	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. laboratory names may contain location information
- Investigator comments may be used to identify a research participant
- Genetic data that would enable a direct trace back to an individual research participant

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 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 for original variable and new anonymized variable)
- The seed utilized for random number generation

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

Appendix 3: A non-real Example Illustrating Removal of Personally Identifiable Information

Center ID	Investigator ID	Investigator Name	Subject ID	Unique Subject ID	Age (years)	AE Start Date	AE End Date	Verbatim Term
00123	279344	Dr Smith	5	TJF00123.0005	57	29DEC2010	27JAN2011	Headache
00123	279344	Dr Smith	2	TJF00123.0002	72	10JAN2011	06APR2011	Nausea
00123	279344	Dr Smith	1	TJF00123.0001	91	25MAR2011	12AUG2011	Cold
05678	333721	Dr Jones	19	TJF05678.0019	85	14OCT2010	20OCT2011	Cold
05678	333721	Dr Jones	4	TJF05678.0004	53	24MAY2011	.	Headache
05678	333721	Dr Jones	23	TJF05678.0023	76	01MAR2011	15MAR2011	Pain

⇩ New Center ID | ⇩ New Investigator ID | ⇩ Remove | ⇩ New Subject ID | ⇩ New Unique Subject ID | ⇩ Replace Age with Ageband | Create new age category | ⇩ Add offset to each date | ⇩ Add offset to each date | ⇩ Remove

Center ID	Investigator ID	Investigator Name	Subject ID	Unique Subject ID	Ageband (years)	Age Category (years)	AE Start Date	AE End Date	Verbatim Term
03145	148227		8754	TJF03145.8754	50-<60	<=89	02FEB2011	03MAR2011	
03145	148227		5681	TJF03145.5681	60-<70	<=89	09NOV2010	03FEB2011	
03145	148227		1475	TJF03145.1475	>=70	>89	03JUL2011	20NOV2011	
90876	687208		1457	TJF90876.1457	>=70	<=89	06JUL2010	12JUL2011	
90876	687208		2214	TJF90876.2214	50-60	<=89	03MAY2011		
90876	687208		2236	TJF90876.2236	>=70	<=89	08MAR2011	22MAR2011	

Alternative approaches include: assigning the same new center ID to low-recruiting centers; removing investigator ID; replacing dates with Relative Study Day.