DH07 Study Anonymisation: From Request to Delivery

Kelly Mewes, Roche Products Ltd, Welwyn Garden City, UK
Sai Jandhyala, Roche Products Ltd, Welwyn Garden City, UK
• “The opinions expressed in this presentation and on the following slides are solely those of the presenter and not necessarily those of Roche. Roche does not guarantee the accuracy or reliability of the information provided herein.”
Our Data Access Model

1. Research Proposal submitted
2. Independent Review Panel
3. Data Sharing Agreement signed (Data Holder(s) & Researcher)
4. De-identified datasets / documents package uploaded
5. Analyses performed within SAS secure website
6. Publication generated
Who are the Requestors?

- All academics from universities

- Evaluating new novel research
  
  - “Assessing models for changes in (melanoma) tumour size over time and how they relate to survival times”
  
  - “Optimized (Pegasys) Treatment of Hepatitis C and Hepatitis B”
  
  - “Identifying immune modulating patterns across diseases from open clinical trial data”

- Multi-sponsor Requests.
Patient Level Data Sharing
Team Operation

• Team
• Remit of group
• Documents and publications
• Governance
Study Data Location and Preparation

- Locating Data
  - Study snapshot to select
  - Electronic submissions
  - Legacy data challenges
  - Other systems for linked documentation
- Provide data in its original state
  - Multiple older data models require preprocessing
  - Able to prepare data packages from multiple areas
Supporting Documentation

- Secure pdf files provided

- Anonymisation Orientation Document
- Details all programming applied
Anonymisation

Why do we need it?

• Maintain **patient confidentiality**
• Direct identifiers need to be removed from the data e.g.
  • Subject name, initials
  • D.O.B
  • Address

• Indirect identifiers in combination could lead to re-identification e.g.
  • Age
  • Geographical location
  • Unusual medical conditions
  • Visible characteristics such as race/ethnicity
Anonymisation Process Flow

**STEP 1 / 2**
Read in and unpack source data to PLDS area. e.g. .xpt or .sas7bdat

**STEP 3**
Perform any pre-processing required to prepare the source data

**STEP 4**
Run SAS macro to anonymise SAS datasets

**STEP 5**
Perform internal independent validation / Code review

**STEP 6**
Final anonymised data created, removing any link back to the original PT

Datasets transferred to Zip Data package for upload to the SAS system
The anonymisation macro
What does it do?

1. Re-coding of Patient and Centre IDs
2. Removing AGE for patients older than 89
3. Dropping variables which could reveal patient identity
4. Converting dates to Study Days and then dropping the dates
5. Applying controlled terminology / grouping certain terms e.g. Race
6. Keeping variables which if dropped would render meaningful analysis impossible

<table>
<thead>
<tr>
<th>Subject ID</th>
<th>New Subject ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>82244</td>
</tr>
<tr>
<td>002</td>
<td>73412</td>
</tr>
<tr>
<td>003</td>
<td>67823</td>
</tr>
<tr>
<td>004</td>
<td>22256</td>
</tr>
<tr>
<td>005</td>
<td>90123</td>
</tr>
</tbody>
</table>

Original Dataset

Mapping Dataset

Anonymised Dataset

Subject ID

New Subject ID
Steps 1 & 2: Prepare Data for Macro

• Copy the study datasets from the original source area to a dedicated anonymisation area

• Unpack datasets and perform pre-processing e.g. date handling

• Data must comply with macro input requirements, such as:
  – Patient and Centre variables must be populated and consistent across selected datasets
  – Dates must be in a compatible SAS format
Step 3:
Run the Anonymisation Macro

%ANON_LIBS2(<INLIBS
,<OUTLIBS
,<DS
,<CENTRE
,<PATIENT
,<DROP
,<MAPLIB
,<EXCLUDE
,<REFDVTAR
,<DEBUG
 = libname(s)>
 = libname(s)>
 = dataset-list|ALL>
 = variable-list>
 = variable-list>
 = filename>
 = libname>
 = dataset-list>
 = variable-list>
 = Y|N>);
Post Macro Execution

• After the macro has been run:

• The anonymised datasets are produced in a dedicated secure location separate to the source

• A PROC contents output of the anonymised datasets is generated for review

• A summary document is produced to aid review which lists the following:
  – Variables that were removed to ensure anonymisation
  – Blank variables that were removed
Steps 4 & 5: Independent Validation

- Visual review of anonymised datasets, with a focus on free text fields, timing variables, and direct identifiers
- Review of PROC contents output
- Review of summary document (orientation document)
- Perform code review of the preliminary and anonymisation programs, final run in Production environment
Step 6:
Data Package Creation

• The data package is created in-line with a template from SAS that contains a standard folder structure as below

  Analysis Ready Datasets
  Documents
  Raw datasets
  Initialisation file

• Parameters of the initialisation file

  Name=RCH-1085-XXXXXX  /*Roche identifier containing Proposal & Protocol ID*/
  Description=Study title  /*As specified in the protocol*/
  SponsorID=RCH-11-2014  /*Default Roche ID value*/
  AppendFiles=Yes  /*Option if you are adding new files*/
Delivery to the SAS CTDT System

- Data delivered via the SAS CTDT MSE
- Secure area for researchers to access the data
- Enables analyses to be performed in SAS/R

Pre-requisites:
- Setup of Study Area(s)
- Creation of Research Area(s)
- User account creation and access
A multi-sponsor scenario

Research Proposal A

Roche STUDY A

Roche STUDY B

GSK STUDY C

Roche STUDY D

Research Proposal B

Roche

GSK

Roche
Future Plans

• Proactive anonymisation
• PhUSE Data De-Identification Standards pilot.
• Automated redaction of free-text variables
• Increased enabling of in house data sharing
Any Questions?

Kelly.mewes@roche.com

Sai.jandhyala@roche.com