PhUSE 2015

Paper 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.

ABSTRACT

Since 2014, Roche has been delivering on external requests for patient-level data. But what happens behind the scenes? Who are the requestors? How does the team operate? This presentation will provide insight including taking you through the steps we follow from receipt of an anonymisation request through to delivery of the final data and documentation package. Topics covered will include identifying the location of study data and the preparation, storage and set-up of anonymised data on our systems and the provision of supporting documentation and how it enhances data access. An overview of our data anonymisation process, how we handle the data, focusing on our anonymisation macro and utility macros and the creation of our anonymisation rules. Our internal QC process will be explained in addition to the steps involved creating the data package and delivery to the SAS CTDT portal and our future plans.

INTRODUCTION

According to Wikipedia data sharing is the practice of making data used for scholarly research available to other investigators. The aim of the Roche policy for patient level data sharing is to maximize the value of clinical trial data to benefit scientific research, increase medical understanding and get the right treatments to the right patients. The aim of our team is to deliver on these commitments to data sharing, whilst protecting patient confidentiality.

Considerations whilst developing the process of data sharing included 'How can researchers request the data?' The process had to deliver a platform to enable researchers to access the list of in scope studies and correct contacts within Roche. 'How can data be shared securely, whilst still enabling researchers to access the data and perform their research?' The process had to strike a balance between the security of the patient level data, to ensure data could not be downloaded and shared outside of the system for example, but still having a high functionality to enable researchers to access and program on an efficient system. This paper considers the solutions we have developed in conjunction with others to address these issues.

PROCESS FOR SHARING PATIENT LEVEL DATA

Roche is part of the multi-sponsor platform clinicalstudydatarequest.com (or CSDR) and the key concepts of the process are as follows:

A research proposal is written and submitted by the researcher, this includes analysis objectives, rationales, statistical analysis plan, researcher affiliations and any conflicts of interest, as well as a list of the required studies. If the researchers' cannot locate the studies they require for their analysis already listed on CSDR, they are able to submit an enquiry to request that the study is listed as available for sharing in order for them to submit their proposal in full. Studies listed on the site have been evaluated as in scope for data sharing via the following criteria and this criteria must be assessed in full before any studies are listed on CSDR.

- Phase II-IV study
- >= 50 patients
- CSR signed off >= 18 months
- FPI >=1Jan99
- Is Roche free from any legal agreement
- Is Roche the Study Sponsor?
- Study used or intended for registrational purpose
- Study Approved in EU and US

Once the research proposal has been submitted the Wellcome Trust assess the research proposal to ensure completeness and following that the study sponsors involved perform a feasibility assessment only to ensure the research proposal meets the following criteria:

- Does the Research Proposal relate to data from clinical studies which are listed on the request site and are the requested datasets and documents available?
- Is the Research Proposal related to the medicine or disease researched in the selected studies?
- Does the research proposal align with the sponsors' requirements?

The access approval of the research proposal in full is assessed by an Independent Review Panel

The study patient level data is anonymized in full, and this is discussed in detail in the following sections.
The study sponsors and researchers sign a legal Data Sharing Agreement to ensure the confidentiality of the patient level data being shared. This covers sections such as not attempting to re-identify the patients and not downloading the data.
Data and the associated supporting documentation is shared via a secure website, which acts as a safe haven for the data. This is hosted by SAS and based on SAS DD, also known as SAS MSE CTDT, this will be discussed in more detail later.
The researchers perform their analysis within the SAS MSE CTDT.
The final step in the process is that the research is published as agreed in the Data Sharing Agreement and a copy is sent to the sponsor for information.

WHO ARE THE REQUESTORS?

So far all of the requestors are academics from universities, some examples as listed on CSDR are listed below and full metrics across all the sponsors listed on CSDR can be found at www.clinicalstudydatarequest.com/Metrics.aspx.
In the UK: University of Dundee and University of Sheffield
In the US: University of California
In the Middle East: The University of Jordan
Most of the research proposals have been evaluating new novel research and not in order to recreate the original study analysis. However, in order for researchers' to understand the data they are handling, researchers may start by recreating the initial study clinical report results before moving on to the new analysis. Some examples of research titles:
"Identifying immune modulating patterns across diseases from open clinical trial data"
"Assessing models for changes in (melanoma) tumour size over time and how they relate to survival times"
"…Bayesian Multivariate Network Meta-Analysis of Ordered Categorical Data (RA endpoints)"
"Optimized (Pegasys) Treatment of Hepatitis C and Hepatitis B"
A large number of our research proposals involve studies from more than one sponsor which illustrates the importance of the multisponsor facility of clinicalstudydatarequest.com and the SAS CTDT portal. The largest multisponsor request involved 6 different sponsors and 15 studies.

HOW DOES THE TEAM OPERATE?

The Patient Level Data Sharing Team ('PLDST') consists of a global lead and a data sharing coordinator, 2 data sharing specialists, 1 in-house contract programmer and 2 contingent resource programmers based in India. We chose to resource using in-house staff as we are the experts on Genentech and Roche data structures and we support others groups at Roche sharing data outside of the company. This includes the sharing of data from project teams with existing collaborators, in addition to sharing via CSDR. As we are sharing data back to 1999 there are multiple data models involved and the experience in handling and manipulating this data is a key factor in ensuring the data is anonymized to a high standard. A separate disclosures group in Roche covers the redaction of CSRs and registry postings to EudraCT and clinicaltrials.gov and another group oversees publications. Close links are maintained with both Disclosures and Publications through internal steering and leadership committees involving all three groups.
Each group has implementation subteams and the authors of this paper are members of the Patient Level Data Sharing Team (PLDST) The remit of PLDST is to:

- Deliver on patient level data requests
- Support project teams internally sharing data externally with collaborators
- Be subject matter experts on data anonymisation and data sharing in general

STUDY DATA LOCATION AND PREPARATION

A vital part of the process is to consider which study snapshot to select from the available repositories, as a large number of the studies handled are oncology trials which have multiple CSR timepoints, e.g. 12 months, 24 months and long term safety follow up. At each snapshot different data may be captured and derived so depending on the data the researchers require, a specific snapshot will be selected.
For studies submitted to the FDA as an electronic submission this is simple as all the documents and data are complete in one package. However if this is not available it can be a challenge to navigate other systems to identify the data snapshot and purpose. This is due to navigating back a number of years across various systems that have been in use due to system migration and company mergers. Once the datasets have been identified, the next step is to locate the linked supporting documentation for that datacut. This again relies on the navigation and searching of the documentation repositories available as there is no simple report to provide this information. There are methods we apply, for example linking the CSR to the data via the CSR table footnotes. We also prepare data packages from multiple areas in order to provide the requested data by our researchers. In this case we take the data from the two storage areas and merge into one study within our anonymisation area and fully detail the original locations and steps taken to merge on our anonymisation orientation document. This has enabled us to share the filing data for a compound but also include a specifically requested endpoint from an extension study that the researchers for their

hypothesis. However, the general principle is that we provide the data in its original state,

As we are handling multiple data models and very old studies we have to preprocess a number of the datasets and variables in order for the standard anonymisation macro to correctly identify the data. Some example of typical processing includes date preprocessing to ensure the dates are correctly handled to be converted into study day.

PROVISION OF SUPPORTING DOCUMENTATION

Supporting documentation we provide within the full data package for our researchers includes

- Dataset specifications
- Annotated Case Report Form
- Protocol
- Reporting and analysis plan
- Redacted Clinical Study Report

We supply all our documentation on secure pdf files this enables researchers to view the documents on the SAS CTDT portal but ensures documents cannot be altered.

The clinical study report most often incorporates the protocol and statistical analysis plan and before sharing these, they are redacted by our disclosures group so all identifiers have been removed such as original patient and centre numbers and all names and contact details.

The most vital piece of supporting documentation we provide is the 'Anonymisation Orientation Document'. This ensures the requestor will be supplied with information describing any changes that have been applied to the data as a result of anonymisation. The document is completed by the responsible data sharing specialist performing the data anonymisation and is independently reviewed by the QC data sharing specialist before being released.

The Anonymisation Orientation document contains a standard section that covers default changes applied to the data by the %anon_libs2 macro. If additional changes are performed (e.g. extra variables dropped, pre-processing, partial date imputation, data cuts etc.) then these changes will also need to be documented in the Anonymisation Orientation document.

The Orientation document consists of the following sections:

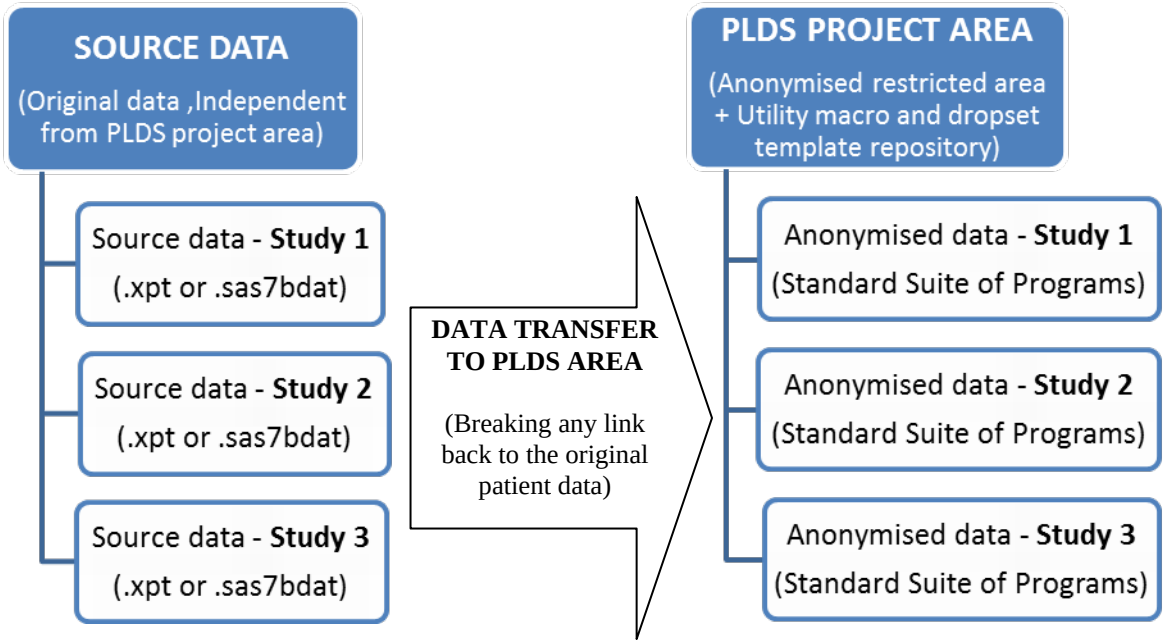
- The study details section is updated manually to reflect the original study number and the research proposal ID taken from CSDR.
- Datasets available and datasets selected to be sent to the requestor, along with the rationale behind the dataset selection. Roche shares the datasets that are identified to contain the minimum variables required for analysis, not each and every dataset within the study. This minimizes the potential risk of re-identification of the subject. The supporting documentation is not updated to reflect this selection, instead all information is kept within so we are transparent as to what is available. If a researcher requests additional data at a later stage, we can share the datasets later if reasonable to do so.
- The anonymisation steps section contains standard text in accordance with the Roche Global Policy on Sharing Clinical Trials Data and requires no alteration. The Roche standards for anonymisation of clinical trial data are in line with a number of available guidance documents from groups such as Phuse Data De-Identification working group and Transcelerate. There are elements included from Hrynaszkiewicz and Norton (2010) which were referenced in the European Medicines Agency (EMA) draft policy on the "Publication and access to clinical-trial data" and HIPAA standards. These standards have been developed in the knowledge that we are sharing in a secure system with named researchers with a legal data sharing agreement in place.
- Modifications that are made to every study as standard – For example within demography patients that are over 89 years old have their AGE variable and Age Category (AGECAT) added.
- Additional study specific modifications applied to datasets in order to anonymize – For example recoding subject to patient for specific studies.
- Table listing every variable grouped by dataset that were removed from the datasets in order to anonymize.
- Table listing every variable grouped by dataset that were removed from the datasets as blank.

THE ANONYMISATION PROCESS AND PRELIMINARY SET UP

Once the supporting documentation is in place, the focus then shifts to the study data. In order to share the data outside of Roche, a process is required to maintain patient confidentiality. Since anonymisation is becoming more frequently required, developing a macro to speed up the process is an important component to saving an organization resources and time. SAS macros help maximize efficiencies by making use of commonly executed processes.

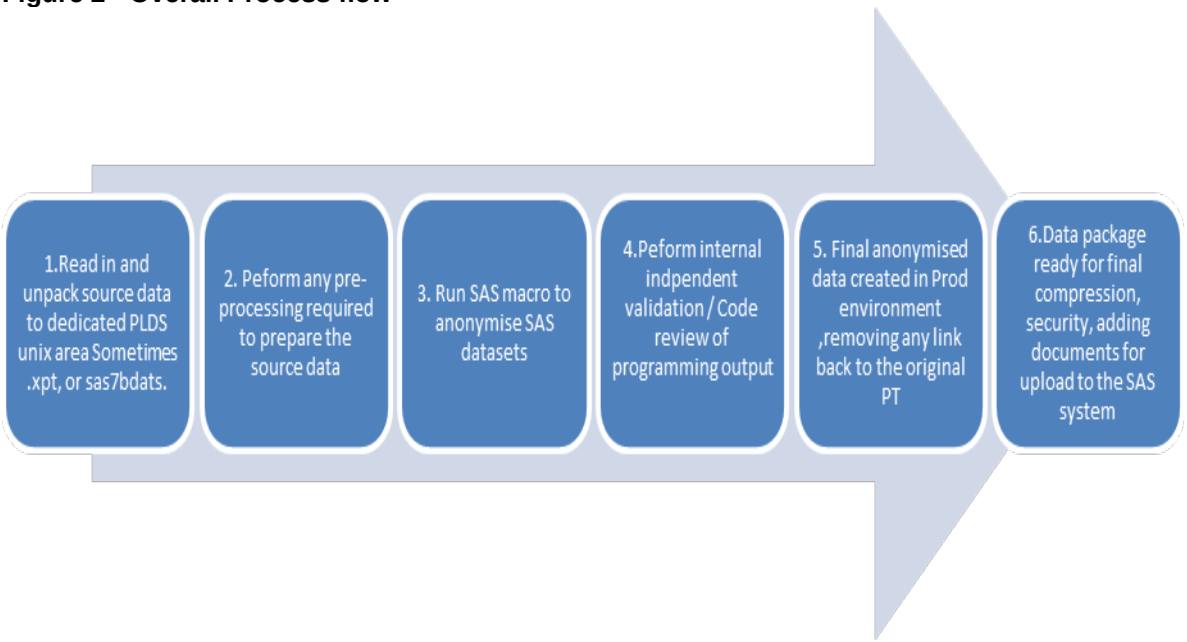
In preparation for executing the macro, the source data is copied from an internal data repository to a dedicated Patient Level Data Sharing (PLDS) project area. This PLDS area is completely separate from the location where the source data is originally stored and analysed. Within the PLDS project area, a standard folder structure for each study is utilised for traceability and audit purposes. A typical PLDS study folder contains a suite of standard programs to support the automated process for anonymising our study data.

Figure 1 – Folder structure and hierarchy



The chart below describes the overall sequence of events in the anonymisation process from data preparation, running the anonymisation macro, QC and producing the final data package to be shared with researcher via the SAS CTDT system.

Figure 2 - Overall Process flow



ANON_LIBS2.SAS

The SAS macro “anon_libs2.sas” was developed by Roche to anonymise the data. The macro is stored and maintained in a centralised macro repository and is updated as required with track revisions. A description of the macro parameters is outlined below:

```
% ANON_LIBS2(<INLIBS=libname(s)>          /*List of libraries you want included in the anonymisation process*/
      ,<OUTLIBS=libname(s)>               /*List of libraries you want to save the anonymised datasets  */
      ,<DS=dataset-list|ALL>              /*Select specific datasets, default is to automatically select ALL datasets */
      ,<CENTRE=variable-list>             /*Names of the centre variables across the datasets */
      ,<PATIENT=variable-list>            /*Names of the patient variables across the datasets */
      ,<DROP=filename>                    /*Name of file containing a list of variables which should be dropped*/
      ,<MAPLIB=libname>                   /*Name of the library where mapping datasets should  be created*/
      ,<EXCLUDE= dataset-list>            /*List of datasets which should be excluded from the anonymisation*/
      ,<REFDTVAR=variable-list>           /*A space-separated list of reference date variables*/
      ,<DEBUG=Y|N>);                      /*Print debugging information when set to Y*/
```

The macro runs via several steps executed in the macro to ensure the data is anonymised, as listed below.

Dropping Variables

Clinical data contains variables which could potentially directly identify a patient, thus these variables must be dropped completely from the anonymised dataset. In a PhUSE 2015 working group paper called "Providing De-identification Standards to CDISC", the authors highlighted important distinctions on the classification of patient identifiers. They defined so called "direct" and "quasi" identifiers, which show the different levels of risk in re-identification of patient data. Direct Identifiers are variables which can be used uniquely to identify a subject e.g. subject name, telephone number and address. It is considered mandatory to either remove or provide a substitute for direct identifiers. Conversely, Quasi Identifiers are defined as background information which in conjunction with other information in the data could potentially reveal a patients identity with a high degree of probability and likelihood of identifying patients versus data utility.

This highlights the importance of classifying variables and when sharing outside an organisation.

Thus, a CSV file is maintained at the PLDS project level containing a list of recommended variables to drop. This list can also be updated as required to ensure it accounts for study specific scenarios. A report of the variables which have been dropped by the macro is created in a summary (".out") file.

Recoding of Patient/Subject ID and Centre Numbers

Each unique patient number is mapped to a number between 20000 and 99999 and each unique centre is mapped to a number between 10000 and 19999.

Additionally, centres with fewer than 10 patients are mapped to 99999 and these are combined with the next smallest centre until there are 10 or more patients assigned to this centre. If the combined centre only has one patient, the country is set to null. Also, if the patient and centre are missing the centre is mapped to missing.

The anonymisation is achieved via a mapping dataset and during development you have the option to reference a permanent copy of the mapping dataset for QC. See example below:

Table 1 – Patient/Subject ID recoding

Original Dataset

Subject ID
0001
0002
0003
0004
0005

Mapping Dataset

Subject ID	New Subject ID
0001	20000
0002	26061
0003	55757
0004	59999
0005	87757

Anonymised Dataset

New Subject ID
20000
26061
55757
59999
87757

It is important to note that the above principle is applied across all datasets. Also, in the production area where the final datasets are created only the anonymised datasets are produced so a patient's data cannot be linked back to the original dataset.

How the macro works

The macro randomly assigns subject IDs, therefore each execution of the macro produces different values for the subject ID as shown below in the code example:

```
proc SQL noprint;
    create table PATIENTS as
      select distinct PATIENT
        from ( %let UNION = ;
               %do I = 1 %to &ANONCNT;
                 %if &&ANONLEV&I = BOTH or &&ANONLEV&I = PT %then %do;
                   &UNION
                   select &&ANONVARP&I as PATIENT
                     from &&ANONILIB&I...&&ANONDS&I
                 %let UNION = union;
                 %end;
               %end; )
        order by PATIENT;
quit;

data ANONP;
    drop ANONNUM;
    do ANONNUM = 20000 to 99999;
      RANDOM = rand('UNIFORM');
      LABEL = put( ANONNUM, 5. );
      output;
    end;
run;

proc SORT data=ANONP out=ANONP( drop=RANDOM );
    by RANDOM;
run;

data _NULL_;
    DSID = open( 'PATIENTS' );
    call symput( 'NOBS', left( put( attrn( DSID, 'nlobs' ), best. )));
    RC = close( DSID );
run;
proc SURVEYSELECT data=ANONP out=MAPP sampsize=&NOBS noprint;
run;

data FORMATP;
    retain TYPE "&PVARTYPE" FMTNAME 'ANONP';
    merge PATIENTS MAPP;
run;
proc FORMAT cntlin=FORMATP( rename=( PATIENT=START ));
run;
```

In the code above, a random sort order for the new subject IDs is created.
The variable LABEL is the new randomly generated Subject ID.
The variable called RANDOM is the sort order that will be used for merging the new subject IDs (numbered 20000 - 99999) with the original subject IDs.
Additionally the PROC SURVEYSELECT procedure, randomly selects which subset of records will be used to merge with the original datasets.
It should also be noted that datasets without centre or patient variables are copied unchanged and the data type of

the centre and patient variable must be consistent (character/numeric) across all selected datasets.

Convert dates to Study Days and then drop the original dates
It is important to consider dates contained within the clinical data increases the risk of of re-identification of a subject. There is much debate within the industry on the best method to accomplish the masking of dates. Some in the industry suggest using an offset on dates to conceal the exact timing of events, however Roche has taken the decision to drop the dates completely and include study day as the timing reference. For SDTM data, the reference start date from the Demographics domain (DM.RFSTDTC) is used as the baseline to re-express all date variables (determined by suffix) as study days. The resulting study day variables will have the same name as the source variable except that the suffix will be replaced with 'DY' e.g. AESTDY will be created from AESTDTC and we follow a similar process for our in house legacy data models.
It should be noted that the baseline reference date can be over-ridden with the REFDTVAR parameter in the macro. This should be provided as a space-separated list of variables in precedence order. For each patient, if one variable contains a missing value, the next in sequence is used. If no non-missing reference date is found a warning is issued. Once the Study Days have been created, all date and date time variables are dropped from all datasets. The exception to this is the birth date variable which is kept on the demography dataset without being re-expressed as study day – the variable AGE suffices. If birth date should be dropped it should be added to the list specified via the DROP parameter as per point 1.

Removing AGE for patients older than 89
Regardless of data model, a new variable called AGECAT is added to the dataset including the AGE variable and takes values <=89 or >89. The AGE is set to missing when AGECAT is greater than 89 or AGE is missing. This is performed as Age is consider critical for analysis and needs to be retained for data utility. It is also required in accordance with the HIPPA privacy rules.

Applying a format to the RACE variable
For non-SDTM studies, RACE will be mapped according to FDA guidelines unless RACE is missing. It should be noted that mapping will not be performed in SDTM studies because this will already have been catered for as controlled terminology.

UTILITY MACROS

There are several utility macros available described below which are stored at the Project level for invocation for any study. These can support the creation of the anonymisation orientation document by auto populating required sections in the document. It should also be noted that the PhUSE De-identification working group are currently working on code sharing and techniques for dropping variables based on the principles of direct and quasi level identifiers with CDISC/SDTM data as mentioned earlier. The utility macros currently help facilitate the process with Roche' own data model and maybe updated to adapt further in the future.

%droplist_freq produces a count of previously dropped variables. This shows how many times each variable was dropped and can be helpful for updating the standard dropset template.

%droplist_find is designed to aid with finding variables to drop from a study that is to be anonymised. It checks all previously dropped variables (by making use of the dropset.csv in all previously anonymised studies) and compares them to the variables in the current study. It produces a list of variables that were dropped in past studies that are also present in your current study. This macro can be invoked as follows **%droplist_find (lib=indata)**

%anon_droplist produces a report of list of variables that were removed to anonymised datasets and generates. The macro syntax is as below with parameter explanation

```
%anon_droplist
    (inlibs   =, /* A space-separated list of input libraries*/
     droplist =, /* Name of file containing variables to be dropped */
     exclude  =  /* Dataset that need to be excluded from the listing (e.g. Comments)
     );
```

%remove_blank_vars produces a report list of blank variables that were removed from each dataset and can be invoked as below:

```
%remove_blank_vars
   (lib     =    /* Input Library */
    exclude =    /* List of datasets to be excluded from processing.
    );              These are usually same as those added to the EXCLUDE parameter of
                    the anonymisation macro */
```

The anon_droplist and the remove_blank_vars macros produce RTF outputs which are appended to the Anonymisation orientation document using the Output Delivery System (ODS) and PROC Report. See example syntax below

```
/*****************************
**    WRITE OUT REPORT      **
*****************************/
ods RTF file="$&__mode/&__proto/reports/blank_vars_&lib..rtf" style=Custom bodytitle;
 title 'The following variables were removed from the datasets as they were all
blank:';
 title2 ' ';
 footnote;
 proc report data=blankvars3 nowd;
  column x dataset concvar ;
  define x / "#"  left;
  define dataset / "Data set name            " width=32 left;
  define concvar / "Variables Dropped"  left flow;
 run;
 ods rtf close;
```

INTERNAL QUALITY CONTROL (QC) PROCESS

For the QC of the process, code review is used rather than dual programming for greater efficiency and to maximise the benefits of a validated macro.
Additionally, using tools in SAS, such as PROC CONTENTS and dictionary tables, the datasets are scanned for the presence of potentially identifying variables. The data is also visually reviewed to ensure all variables that could identify a patient have been removed, including any special cases such as study specific nuances.
Free text variables pose a large risk to patient privacy, because they may contain language that identifies a patient. Therefore, we look at variable labels that may contain keywords such as 'Page' , 'Other' , 'Reason', 'Comment', 'Specify' ,'Text' , 'LOT' , 'Initial' , 'Inv' , 'Name' , 'Route', 'Unit', 'Dose' because these labels have been identified as potential free-text fields. Additionally, the study day variables are checked to confirm that the dates have followed the standard imputation requirements as applicable.
After all checks have been completed and issues resolved, all the anonymisation programs are promoted to the production environment and execute a batch run of the anonymisation programs checking the final anonymised datasets are populated as expected in the output library. It should be noted that a review the Anonymisation Orientation document is performed to ensure it accurately reflects ALL the processing performed on the data.

DATA PACKAGE CREATION AND DELIVERY TO THE SAS CTDT MSE SYSTEM

Now that the data has been anonymised and passed quality control we need to transfer the anonymised datasets along with supporting documentation to a zipped data sharing package and upload to the SAS Multi Sponsor Environment via the SAS secure Clinical Trials Data Transparency Portal.
It should be noted that the data package is created in-line with a template provided by SAS that contains a standard folder structure split into the following sections:

- Analysis Ready Datasets
- Documents
- Raw datasets
- Initialisation file

An .ini file contains information so that the MSE will process and create the Research Study Area in the system. In the template, the .ini file will be as below:

```
Name=RCH-1085-XXXXXX     /*List of datasets to be excluded from processing.
                           These are usually same as those added to the EXCLUDE
                           parameter of the anonymisation macro */
Description=Study title  /*As specified in the protocol*/
SponsorID=RCH-11-2014    /*Default value*/
AppendFiles=Yes          /*Can be added as an option if you are adding new files*/
```

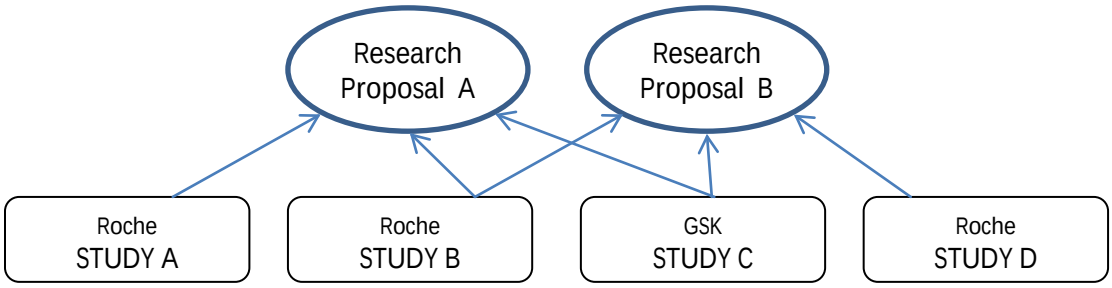
In order for the researcher to access their requested data and documents three main steps then follow:

- Setup of Study Area(s)
- Creation of Research Area(s)
- User account creation and access

The first step requires for the data package(s) to be uploaded via the SAS CTDT portal to the SAS study area repository. This contains all the data uploaded by a sponsor. An email confirmation from the system notifies you that this stage has been successful. Study data uploads can either be performed via a Secure File Transfer Protocol (SFTP) link from our UNIX platform or directly from the CTDT point and click interface. The SFTP alternative method can be useful when uploading large datasets (exceeding 100mb).

The next step involves setting up the Research Areas are where the researchers develop their programs and access the data. It should be noted that a study can be accessed by more than one research area. See Illustration below:

Figure 3 – A Multi-sponsor Scenario



The illustration above displays two research areas Research Proposal A and Research Proposal B which are linked with various studies. If a researcher requests a granted proposal, e.g. Research Proposal A, they will be able to access the data from Study A, B and C but not D. Similarly, a researcher with a granted proposal equivalent to Reasearch Proposal B will be able to access the data from Study B, C and the final step involves adding users to the research area and this is done by sending a request email to SAS outlining the researcher's names and details. A Data Access Form is completed by the researcher to help facilitate this and this step is normally performed at the time of requesting SAS to setup the required Research area.
Now that the anonymised data and documents have been successfully uploaded and the required research area has been setup with associated users a final QC can be performed.
It should be noted in the above multisponsor example, Roche would load the data for studies A,B,D and GSK would load the data for study C. An individual sponsor cannot see data belonging to another but reports are available to check setup and access is correct.
The data package and supporting documentation are now available for researcher via to access and conduct their research.

FUTURE PLANS

Our future plans include moving towards proactive anonymisation of studies that we have listed on CSDR. We are currently considering the role of formal quantification of risk of patient re-identification, (if any, given the context in which we share data in a controlled manner), in the current processes we follow within our study anonymisation data flow. Roche continues to play an active role in the Phuse Data De-Identification Standards working group and are running a pilot on our internal SDTM data which we hope to feed back to the group soon. We are looking to further to develop our tools for better automated anonymisation of free-text in our dataset records, which enables free text

variables that are of high scientific value to be kept within datasets but have all potential identifiers, such as dates and names, replaced with *redacted* via an automated method rather than manual review.

CONCLUSION

The objective of this paper has been to provide an overview of the process Roche follows from the point of receiving the initial request up to the final delivery of a Data Sharing Package to the requestor. Consideration of current guidance and common practices within our industry has been provided focusing on the key principles of preserving patient privacy when sharing clinical trial data outside of an organization.

From the last year, it can be concluded that there is a varied range of requestors, looking to perform novel research which could not only yield huge benefits to patients but also maximizes the wealth of existing clinical trial data already available.

In terms of our anonymisation process, future enhancements are constantly being considered and it is envisaged that workflows will adapt and be tailored to cater for more automation whilst still adhering to the current anonymisation and CDISC industry standards.

Finally, cross company collaboration and raising industry awareness should be recognized in ensuring that future study anonymisation and data de-identification undertaken by pharmaceutical companies are aligned and consistent in their approach, Working groups from PhUSE play an important role here and as the standards mature it is expected that increased alignment will be seen across this emerging, developing and exciting landscape.

ACKNOWLEDGMENTS

Thanks to Katherine Tucker for her review and input.

RECOMMENDED READING AND REFERENCES

1. Clinical Study Data Request.com. https://www.clinicalstudydatarequest.com/.
2. TransCelerate Biopharma Inc. http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/Data-Anonymization-Paper-FINAL-5.18.15.pdf.
3. Pharmaceutical Users Software Exchange. Data Transparency. http://www.phuse.eu/Data_Transparency.aspx.
4. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers.
5. All Trials http://www.alltrials.net.
6. Federation of Pharmaceutical Industries and Associates (EFPIA) – PhRMA. Principles for Responsible Clinical Trial Data Sharing: Our Commitment to patients and researchers. http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.
7. Shostak J. De-Identification of Clinical Trials Data Demystified. http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr02.pdf.
8. International Pharmaceutical Privacy Consortium. IPPC White Paper on Anonymisation of Clinical Trial Datasets. Oct 2014. http://pharmaprivacy.org/activities/ippc-white-paper-on-anonymisation-of-clinical-trial-data-sets.
9. Gibson B. Multi-Sponsor Data Transparency: A Group Approach to Sharing. 2014. http://www.phusewiki.org/docs/Conference%202014%20TT%20Papers/TT04.pdf.
10. Ferran JM, Lanoue J. PhUSE De-Identification Working Group: Providing de-identification standards to CDISC data models 2015 www.pharmasug.org/proceedings/2015/DS/PharmaSUG-2015-DS10.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kelly Mewes
Roche Products Ltd
Hexagon Place, 6 Falcon Way, Shire Park
Welwyn Garden City, Hertfordshire, AL7 1TW
Work Phone: 01707 366206
Email: kelly.mewes@roche.com

Sai Jandhyala
Roche Products Ltd
Hexagon Place, 6 Falcon Way, Shire Park
Welwyn Garden City, Hertfordshire, AL7 1TW
Work Phone: 01707 366106

Email: sai.jandhyala@roche.com