PhUSE US Connect 2018

Paper TT14

Clinical Trial Datasets (CDISC - SDTM/ADaM) Using R

Prasanna Murugesan, AstraZeneca, Gaithersburg, USA

ABSTRACT

Open source statistical software R is being used in several industries for data analysis and data visualizations to provide meaningful insights. Although, R has been used in exploratory analysis in Pharma/Biotech industry for a long time, it has not been used for creating and analyzing clinical trial datasets (SDTM/ADaM). Traditionally, SAS® has been used to generate clinical trial datasets. In this paper we will explore the technical feasibility of generating SDTM/ADaM datasets using base, dplyr, haven, lubridate and bridge packages of R.

INTRODUCTION

Modern data capturing techniques using mobile apps, smart devices has given the ability to collect huge amount of data from end users. Companies are looking for software tools to analyze the collected data and interpret meaningful insights using the collected data.

In the recent past, R is being used in several industries for data analysis and data visualizations to provide such insights. Although, R has been used in exploratory analysis in Pharma/Biotech industry for a long time, it has not been used for creating/analyzing clinical trial data sets.

In a highly regulated Pharma/Biotech industry, we are advised to use validated systems that have gone through the rigor of Software Development Life Cycle(SDLC). Although, we use open source tools to check data quality of clinical trial data sets, open source computing software like R has not been used to analyze clinical trial data especially for regulatory submission purposes. While R does have some packages written by users that have not gone through the formal SDLC process, this paper will see if it is technically possible to use R and to generate SDTM/ADaM data sets and leverage the functionality provided by a compelling cost-effective software available to us.

R-PACKAGES AND SAS® EQUIVALENTS:

In this paper, we will explore if R can be used for clinical trial data manipulation and creation of CDISC: SDTM/ADaM data sets. The following R packages were used during this evaluation.

- base : Base R functions
- dplyr : designed for data transformation
- lubridate : Date functions
- haven : Loads SAS® data sets
- bridge: generates infile SAS® program and CSV file for a data set

The following table compares some of the frequently used SAS® syntaxes with its corresponding R equivalents.

SAS® syntax	R:dplyr equivalent
keep/drop	select
all data step derivations	mutate

by statement	group_by
if/where statement	filter
proc sort	arrange
statistical procedures (proc means/tabulate)	summarise
transpose	spread/gather
merge	left/right/inner/full joins
set by rows/columns	bind_cols/bind_rows

R:DPLYR PROGRAMMING

Being a SAS® programmer, I started to work on the program with very simple approach as we take during SAS® programming. The following topics will discuss some of the key segments that were used in the R program. For easy readability, please read the symbol "%>%" as "and then". It means that we process some data "and then" pass that data to the next step.

LIBNAME SETUP

```
sdtm <- "//product/study/analysis/data/sdtm" # assign dir to object named sdtm
out  <- "//product/study/analysis/data/adam"
```

READ FILES

```
dm   <- read_sas(file.path(sdtm,"dm.sas7bdat")) # read sas file as a data frame
ds   <- read_sas(file.path(sdtm,"ds.sas7bdat"))
sv   <- read_sas(file.path(sdtm,"sv.sas7bdat"))
suppsv <- read_sas(file.path(sdtm,"suppsv.sas7bdat"))
```

MERGING PARENT AND SUPPLEMENTAL DATA SET

```
suppds_     <-   suppds %>%                # "%>%" read as "and then"
                 mutate(idvarval_ = as.numeric(idvarval)) %>%
                 select(usubjid,idvarval_,qnam,qval) %>%
                 spread(.,qnam,qval)

ds_all      <-   left_join(ds, suppds_,
                           by = c("usubjid"="usubjid",   "dsseq"="idvarval_"))
```

BASELINE FLAG/CHANGE FROM BASELINE DERIVATION

```
advs <-   advs_ %>%
          group_by(subjid,paramcd) %>%
          arrange(subjid,paramcd)  %>%
          mutate (  base = aval[visitnum==1],
          ablfl = ifelse(visitnum == 1,"y",na),
          chg   = ifelse ( !is.na(aval) & !is.na(base),aval-base,na),
          pchg  = ifelse ( !is.na(aval) & !is.na(base),((aval-  base)/base)*100,na)
```

SUBJECT LEVEL DERIVATION

```
adsl <-   dm %>%
    select(studyid, subjid, age, sex, height, weight, race, scrfl) %>%
    mutate(bmi = (weight*703)/height^2 ) %>%
    filter(scrfl == “Y”) %>%
    select(-scrfl) %>%
    arrange(studyid, subjid)
```

Using the techniques shown above DM, ADSL and ADLB datasets are generated. As you notice, R-dplyr allows the user to merge, transform, create new variables in one step without a need to sort the dataset in every single step. R-haven/bridge package can be used to convert R output to SAS7BDAT file. R-haven/bridge package will convert the final file to a csv file and a provide SAS® code to convert the csv file to SAS7BDAT file when executed in SAS® environment.

CONCLUSION

The ability for R to create clinical trial datasets can be looked as a potential alternative, cost effective way compared to existing methods. Also, R code can be easily integrated with data visualization tools like R-Shiny.

Some of the challenges faced during R programming are R’s inability to provide a detailed log and the inability to add labels to variables/datasets. You also need to stay up to date with all the change to packages and analyze the impact of such changes to existing code.

REFERENCES

SAS® AND R - STOP CHOOSING, START COMBINING AND GET BENEFITS!
https://www.pharmasug.org/proceedings/2016/QT/PharmaSUG-2016-QT14.pdf

SAS® AND R PLAYING NICE TOGETHER
https://www.pharmasug.org/proceedings/2017/PO/PharmaSUG-2017-PO22.pdf

ADDITIONAL READING

http://dplyr.tidyverse.org/

https://rpubs.com/bradleyboehmke/data_wrangling

CONTACT INFORMATION

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

Author Name: Prasanna Murugesan
Company: AstraZeneca
City / Postcode: Gaithersburg, USA
Email: Prasanna.murugesan@astrazeneca.com