PharmaSUG 2018 - Paper SI-12

Exploring Use of R for Clinical Trials

Kalpesh Prajapati, Parveen Kumar, GCE Solutions

ABSTRACT

SAS is leader for data analysis in health care industry being accepted by regulatory bodies worldwide. R is one of the open source software used in academic and other research firms. As regulatory bodies never endorse any particular software, R can revolutionize the analysis part of clinical trials.

INTRODUCTION

Since last few decades clinical industry is evolving, and it requires standardization all over the world. SAS played major role in analysis of data for submission. SAS is obvious leader for analysis in health care industry as it is recognized and accepted by regulatory bodies across the world. Industry folks find it easy to learn but expensive one when it comes to individual use. On the other hand, there are software available as open source. R is one of those freeware which is now preferred by academics and other research firms as it is cost effective and easily available compared to SAS. Though SAS is easy to learn and provides simpler coding options, R has stepwise learning depending upon programming language. Many education institute provide opportunity to learn R as part of their curriculum. R has better graphical capabilities too when it comes to data presentation. Also, regulatory bodies never endorse any software to use for submission trial. Any software compliant to regulations and requirements is acceptable.

This paper will discuss about the status of use of R in Pharmaceutical industry in perspective of regulatory submissions and compare its use with SAS. It will also be discussed that whether R can revolutionize the analysis part of clinical trials in industry.

DISCLAIMER

The scope of this paper is to present the opinions and suggestions of the author. The interpretations of standards and procedures contained in this paper are those of the author and they do not represent the position of their employer.

ABOUT R

R is a high-level, interpreted programming language-based software which is known for statistical analysis and graphical reporting. It allows data analysis, manipulation, graphical presentation in easy and effective way. The use of R increased exponentially over last decade considering its easy access and popularity. R (the successor of S) was created in 1992 by Ross Ihaka and Robert Gentleman at University of Auckland. Now it is developed by the R Development Core Team.

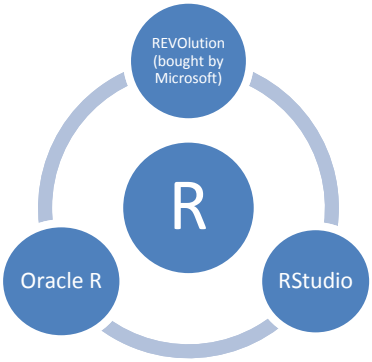
R is easily available under GNU General Public License which is being used for statistical analysis by data analyst across corporates and academics. Now, many of the researcher and academician using it widely for their personal and professional data analysis.

I would advise to know few things about R, before reading this paper. This is not an attempt to provide guidance on R programing. Some basic information about R programming would be advantage.

R works on different operating systems such as Windows, Unix, and MacOS. Programming language in R consists of interpretable syntax, functions, procedures. Base R includes but not limited to base, compiler, datasets, graphics, grDevices, grid, methods, parallel, splines, stats, stats4, tcltk, tools, utils.

R provides a wide variety of statistical tests such as linear and nonlinear modelling, regression analysis, classical statistical tests such as analysis of variance, time-series analysis along with wide graphical techniques. Various R packages exists on CRAN (Comprehensive R Archive Network) consisting of several functions. All functions need not to be understood in order to perform analysis tasks. You can understand basic and essential R functions to get the grip on software.

R offers continued development for new methodology, training to beginners and intermediates for statistical computing. R provides support to individuals, institutions and commercial enterprises. There is misconception that R is not validated and cannot be used for clinical trials. But, it is organization's responsibility to work on standard operating procedures(SOPs) for installation, validation and utilization. Else, as shown below, software companies have come up to support with use of R. These companies provide R in modified interface free as well as commercial.



R development core team including academic, non-profit and institutions from various statistical disciplines develop and maintain R. This team works on development and testing methodologies for more accurate, reliable and consistent performance. Source codes in R, are managed by version-controlled repository. These repositories are access controlled. There are validation test/process for source code which are maintained and upgraded by core team.

R PACKAGES

R homepage provides information on features, basic understanding and add-on packages. Currently, R supports more than 12,000 packages in CRAN repository. Current and historical versions of R and these packages are available from the main CRAN server. Apart from some default packages available in R, more than 6000 packages available online depending upon techniques as Bayesian, Survival, timeseries, Adaptive designs, experimental designs, etc.

COMPARISON OF R VS SAS

SAS and R are different primarily considering data management, analysis with wide-ranging options available.

COSTING AND SUPPORT

As R is easily available open source software and SAS is licensed software, there are number of differences including software validations, sales, and support. SAS has a team of paid qualified computer scientist, programmer, subject matter experts and a biostatistician working towards betterment of codes and its implementation along with acceptance of far advance techniques in terms of statistical analysis. For R, number of users from community of researchers and methodologist supported, developed and implemented functions for R. But being flexible, code quality may be discussed and improved by others. Also, advancement in codes or functions may lack compatibility with older versions, improper revised documentation to guide further, and inexistence of guaranteed support. For SAS, there are books and SAS specific literatures clearly written and equally reliable such as categorical Data Analysis using SAS, Survival Analysis using SAS.

Private sectors including enterprises prefer SAS considering long history of ease of doing businesses, evolved interfaces and extended functionalities over the time. Since long back, effective, standard and statistically strong quality codes are result of hours of paid resources. SAS offers great customer service and support. R may need time to conceivably match the level as there is no paid support for free version. However, as discussed in earlier sections, there are groups supporting R commercially, such as Revolution analytics, RStudio(commercial).

TECHNICAL

SAS operates at observation level whereas R offers vectors-based coding. SAS is generally not case sensitive whereas R is case sensitive. R has come a long way and analysts are aware of it. R is very suited for medical research and hence, many of the top pharmaceutical companies building strengths within the team.

Many of the required functions from statistical analysis perspective are available in R as in cores of SAS. If one is familiar with SAS, R is as easy and interesting as SAS. R can also be accessed conveniently from SAS via IML module. For beginner with statistics background and not much of programming exposure, it can be slightly challenging considering moderately available literatures compared to SAS learning. Assumption that R has limits to handle large data is not quite true though. Revolution Analytics maintains a proprietary version of R and they claim to have solved the in-memory limitations of R. Many users prefer R mainly for its graphical facilities.

There are ways in which user can take advantage of both SAS and R without totally switching from SAS to R or vice versa. Currently, R and SAS are often used parallelly may be to provide data to R user or to receive data from R.

SOME EXAMPLES

Few of the methods which needs to be programmed in SAS are easily available in R packages. For example, confidence interval for the difference of two proportions using EXACT RISKDIFF within PROC FREQ or other exact 95% CI using METHOD=FMSCORE can be produced in SAS. R has ExactCIdiff package available from CRAN. This package contains 1) PairedCI() for calculating lower one-sided, upper one-sided and two-sided confidence intervals for the difference of two paired proportions and 2) BinomCI() is for the difference of two independent proportions. Exact intervals for the difference of two paired proportions can be calculated using PairedCI() function of ExactCIdiff whereas, SAS does not compute exact intervals for the difference of two paired proportions at all.

```
PairedCI(n12, t, n21, conf.level, CItype, precision, grid.one, grid.two)

BinomCI(n1, n2, x, y, conf.level, CItype, precision, grid.one, grid.two)
```

Here, n12, n21, t(n11+n22) and n1, n2, x and y are the observations from the experiment, conf.level is the confidence coefficient of the interval, CItype to get upper one-sided or a two-sided interval and precision of the confidence interval in decimal places. The values of grid.one and grid.two are the number of grid points in the two-step approach to search the global maximum of the tail probability in the first step and second step respectively.

SAS has advance ODS system for producing rtf and pdf outputs. R offers advance reporting as ReporteRs for Word and powerpoint document generation.

While dealing with huge data, SAS offers SAS/ACCESS module which is useful for clinical research. SAS Transport files (XPT) are mode to submit data to regulatory. These XPT can be read into R with standard read.xport function and exported from R with the write.xport function in the SASxport package. Details are covered in Data Exchange between R and SAS.

DATA EXCHANGE BETWEEN R AND SAS

SAS to R

Using SAS IML procedure, exchange of data becomes easier. It allows R advantages within SAS environment. Using SUBMIT/ENDSUBMIT statements in IML procedure and a few CALL routines, you can create R data frames from SAS data sets and execute R statements within your SAS programs and take advantage of both SAS and R software.

For e.g. SAS/GRAPH module is for graphical presentation in SAS to produce efficient graphs. Similar way, ggplot2 is a plotting system for R which helps to produce quality graphics.

Below is and extract of SAS code to prepare graph using R in SAS interface. As discussed above, ggplot2 is known package and can be installed once and loaded every time with new R session.

```
proc iml;
   run ExportDatasetToR("data");
   submit/R;
   attach(data)
   install.packages("ggplot2")
   library(ggplot2)
endsubmit;
quit;
```

Below table shows modules for transferring SAS to R.

Table 11.1 Transferring from a SAS Source to an R Destination

Method or Module	SAS Source	R Destination
ExportDataSetToR	SAS data set	R data frame
ExportMatrixToR	SAS/IML matrix	R matrix
DataObject.ExportToR	DataObject	R data frame

http://support.sas.com/documentation/cdl/en/imlsstat/63545/HTML/default/viewer.htm#imlsstat_statr_sect004.htm

R to SAS

ASCII Text Files : The WRITE.FOREIGN() function in the FOREIGN library writes a text file along with the SAS program to read it in.

```
library(foreign)
write.foreign(data,"study/xyz.txt","study/xyz.sas",
package="SAS")
```

SAS Transport Files : The WRITE.EXPORT() function in the SASXPORT library creates an xpt file.

```
library(SASxport)
write.xport(data, file="study/xyz.xpt")
```

SAS IML : Reading data from R in SAS IML will be same as to data from SAS to R.

Below table shows modules for transferring R to SAS.

Table 11.2 Transferring from an R Source to a SAS Destination

Method or Module	R Source	SAS Destination
DataObject.AddVarFromR	R expression	DataObject variable
DataObject.CreateFromR	R expression	DataObject
ImportDataSetFromR	R expression	SAS data set
ImportMatrixFromR	R expression	SAS/IML matrix

http://support.sas.com/documentation/cdl/en/imlsstat/63545/HTML/default/viewer.htm#imlsstat_statr_sect005.htm

This is the easiest way of working with SAS and R together, of course with a caution.

ACCEPTANCE TO REGULATORY

Certain guidance on regulatory requirements for electronic records and electronic signatures as per CRF part 11 issued by FDA to be followed. One may have question about validation of R to consider its use in clinical trials.

Guidance from FDA, Computerized system used in clinical trials considers software validation as *"confirmation by examination and provision of objective evidence that software specifications conform to user needs and intended uses, and that the particular requirements implemented through software can be consistently fulfilled.*" Validation definitions as per FDA is *"Establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes*."

Regulatory bodies never endorsed on use of any specific software for analysis and reporting of clinical trial submission. So, they never restricted use of any open source software. SAS, R or any other compatible software used for data analysis should comply with required regulations. Results should be reproducible and independent of the software used to derive those.

So, validation is required to guarantee quality and reliability of a process or product thoroughly. R confirms that it supports appropriate regulatory requirements for validated systems when used in qualified fashion.

Here, R provides "Regulatory Compliance and Validation Issues: A Guidance Document for the Use of R in Regulated Clinical Trial Environments" which talks about regulations specified by regulatory bodies including largely for USFDA (The United States Food and Drug Administration) and ICH(International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals in Human Use) and also for EMEA (European Medicines Agency) and PMDA (the Japanese Pharmaceuticals and Medical Devices Agency) but not extensively as for FDA and ICH. It is not new that FDA is accepting reports created with R, it is there since long time.

ORGANIZATIONS USING/USED R

People understands that SAS is for business and R is for individual research or academics. When one dig into research, corporate world including big pioneers of the time are referring R at some point. Some of the organizations are using it for daily operations too. AstraZeneca use R and RStudio data science tool to work with medical data. As per Paul Metcalf, "With R, RStudio, and Shiny we've created a robust tool chain for routine tasks and enabled reproducible research."

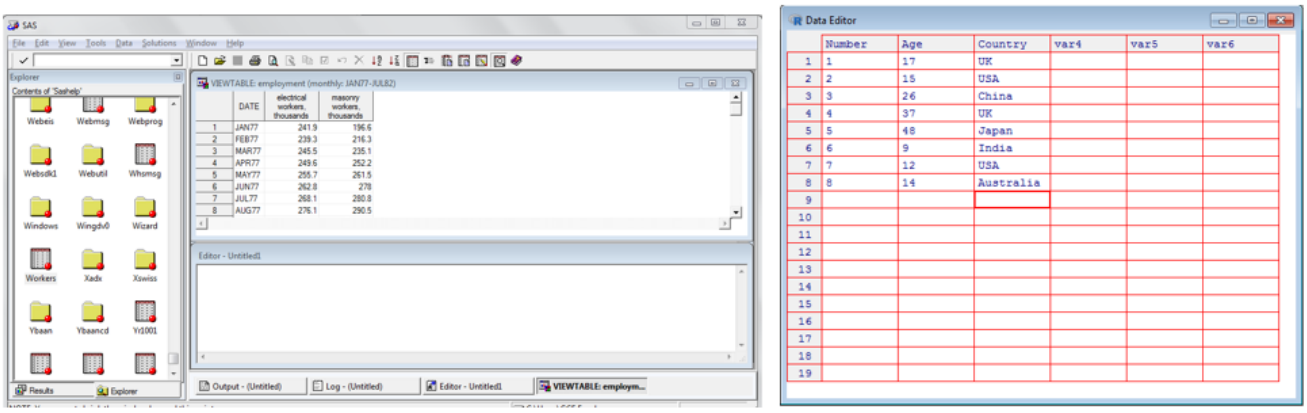
As per Zach Schleien, ITLDP Analyst, Business Technology Leader at Johnson & Johnson, The Janssen Global Epidemiology department has built a tool using R and Shiny to conduct network meta-analyses using data from ClinicalTrials.gov. Other companies like Accenture, Nestle, Novartis, Medtronics, Lilly, considered use of R at certain points.

LEGACY DATA

SAS is a part of pharmaceutical industry since early ages i.e more than 30 years. Analysis done in past or codes prepared are compatible/readable in SAS only. It would take few years of dedicated resources to convert all those legacy codes to R. This calls for investment as well or else legacy data would not be available for further reference or reproducing results. This is also one of the reasons that R is not capturing the pharmaceutical market very fast.

DATA VISUALIZATION AND MANAGEMENT

Preparing datasets which are analysis friendly needs lots of peek into intermittent datasets to evaluate the programming ongoing for conversion. SAS has user friendly interface to inspect datasets stored in directory. Multiple commands for filtering datasets are available on the interface itself. On the contrary, it is slightly difficult to inspect datasets on R interface.



Screenshot of the Viewtable in SAS and R data editor

CONCLUSION

There are many supplementary matters which could be discussed for R and SAS. However, comparisons only at some functions are discussed here. As primary goal is to determine safety and efficacy, it should be independent of software used for analysis. But, results should be reproducible irrespective of software.

SAS is leader for data analysis in health care industry being accepted by regulatory bodies worldwide. R is one of the open source software used in academics and other research firms. Regulatory do not prevent use of R considering validation and documentation required. Considering that software has to be compliant to regulations and requirements of regulatory, R can revolutionize the analysis part of clinical trials.

One can argue whether SAS or R has better programming language or specific procedure is better in SAS or R. The discussion should be whether one should expect significant changes in the future as both R and SAS software are not new anymore. SAS is very well known to clinical research industry being acclaimed statistical software for data analysis since more than three decades. The advantage of SAS is availability of various case studies, easily available documentation, readily available codes with examples. As most of the statisticians and programmers are good at programming in SAS, they select software depending upon his or her knowledge about statistics and programming skills. If one is good at both, not so hard to accept any programming software. If you already know statistical packages, time you need to learn R depends upon your interest of learning specific function. This may help to make you more interested in learning other packages once you know what R offers. So, it will be appealing solutions to use both, SAS and R to improve your data analysis and presentation abilities.

REFERENCES

Ken Kleinman and Nicholas J. Horton, 2014. *SAS and R: Data Management, Statistical Analysis, and Graphics*, Second Edition, CRC Press, Taylor & Francis Group

Robert A. Muenchen, 2011. *R for SAS and SPSS Users*, Second Edition, Springer.

H. Wickham. ggplot2, 2009.*Elegant Graphics for Data Analysis*. Springer, New York.

Guogen Shan and Weizhen Wang, December 2013. “ExactCIdiff: An R Package for Computing Exact Confidence Intervals for the Difference of Two Proportions” The R Journal Vol. 5/2.

Matthew Cohen, 2012, *“SAS® and R Working Together”,* NESUG

The R Foundation for Statistical Computing. *R: Regulatory Compliance and Validation Issues A Guidance Document for the Use of R in Regulated Clinical Trial Environments,* December 15, 2014. www.r-project.org/doc/R-FDA.pdf

The R Foundation for Statistical Computing. *R: Software Development Life Cycle A Description of R’s*

Development, Testing, Release and Maintenance Processes, December 15, 2014 https://www.r-project.org/doc/R-SDLC.pdf

"Is R suitable enough for biostatisticians involved in Clinical Research & Evidence-Based Medicine?" 15 Jun 2015, http://www.r-clinical-research.com/

Diana Bulaienko, 2016 *"SAS® and R - stop choosing, start combining and get benefits!",* PharmaSUG 2016.

For ReporteRs package, https://cran.r-project.org/web/packages/ReporteRs/index.html.

For ExactCIdiff package, https://cran.r-project.org/package=ExactCIdiff

For user of R: RStudio, https://www.rstudio.com/resources/customer-spotlight/astra_zeneca/

For user of R: RStudio, https://www.rstudio.com/resources/customer-spotlight/janssen-story/

Guidance for Industry: Part 11, Electronic Records; Electronic Signature - Scope and Application, August 2003.

Guidance for Industry Computerized Systems Used in Clinical Investigations, May 2007

ACKNOWLEDGMENTS

We would like to thank Rajan Sareen at GCE solutions, for providing required guidance and supportive documents to write this paper.

CONTACT INFORMATION

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

Kalpesh Prajapati
GCE Solutions
+31 633692774
kalpesh.prajapati@gcesolutions.com
www.gcesolutions.com

Parveen Kumar
GCE Solutions
+91 9871819209
parveen.kumar@gcesolutions.com
www.gcesolutions.com