PhUSE US Connect 2018

DH03

Cross Study Analysis of SDTM Data Using R

Mike Collinson, Todd Cornelius, and Greg Jones, Oracle Health Sciences;
Reading, UK and Bedford, USA.

ABSTRACT
The growth in the range of inter-connected devices across healthcare represents an exponential growth in the volume of data collected in ever more elaborate Clinical Trials. This growth in the volume of data presents new challenges for Clinical Data Scientists and requires new solutions and new tools for cross-study data analysis.

To meet these demands, Clinical Data Scientists are increasingly choosing open source solutions to leverage the active open source communities of experienced developers and statisticians. The R scripting language is increasingly popular in the biostatistics and statistical programming fields and supports predictive analytics, big data analysis, and offers the potential to leverage Machine Learning and Artificial Intelligence.

Regulators already accept R for statistical analysis and the requirement for skills in R is growing faster than other competing tools. This presentation will look at the use of R and related technologies in cross study data analysis using SDTM data.

INTRODUCTION
Clinical research is experiencing a revolution with a huge range of connected devices growing in popularity, with wearable and implantable devices across healthcare, fitness tracking and diet. Pharmaceutical companies sponsoring trials are incorporating these devices into ever more elaborate clinical trials, generating ever larger datasets, while sifting through social media streams and their own big data sources. It is now easier than ever before to store, manage and query ever increasing datasets.

The growth in the range of inter-connected devices across healthcare represents an exponential growth in the volume of data collected in ever more elaborate Clinical Trials. This growth in the volume of data presents new challenges for Clinical Data Scientists and requires new solutions and new tools for cross-study data analysis.

To meet these demands, Clinical Data Scientists are increasingly choosing open source solutions to leverage the active open source communities of experienced developers and statisticians. The R scripting language is ever more popular in the biostatistics and statistical programming fields and supports predictive analytics, big data analysis, and offers the potential to leverage Machine Learning and Artificial Intelligence.

Regulators already accept R for statistical analysis and the requirement for skills in R is growing faster than other competing tools. This presentation will look at the use of R and related technologies in cross study data analysis using SDTM data.

R – OPEN, EXTENSIBLE, SCALABLE, AVAILABLE
Today, written words and numbers are everywhere, unending and ever - changing. In this world of infinite variety, visuals are still the best way to tell a story. In fact, visualization is more important than ever, because with all the information that's available, it's getting harder and harder to sift through the clutter to understand what's valuable. Today, visualizations are the best way to filter out the noise and see the signals.

R is a statistical and visual language used by a growing number of data analysts inside corporations and academia, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America and Shell use R.

It is also free. Open-source software is free for anyone to use and modify so statisticians, engineers and data scientists can improve the software's code or write variations for specific tasks. Packages written for R add advanced algorithms, richly coloured and textured graphs and mining techniques to dig deeper into databases. Pharma companies have created customized packages for R to let scientists manipulate their own data during nonclinical drug studies rather than send the information off to a statistician, and R continues to grow both in popularity with Clinical Data Scientists, and in completeness:

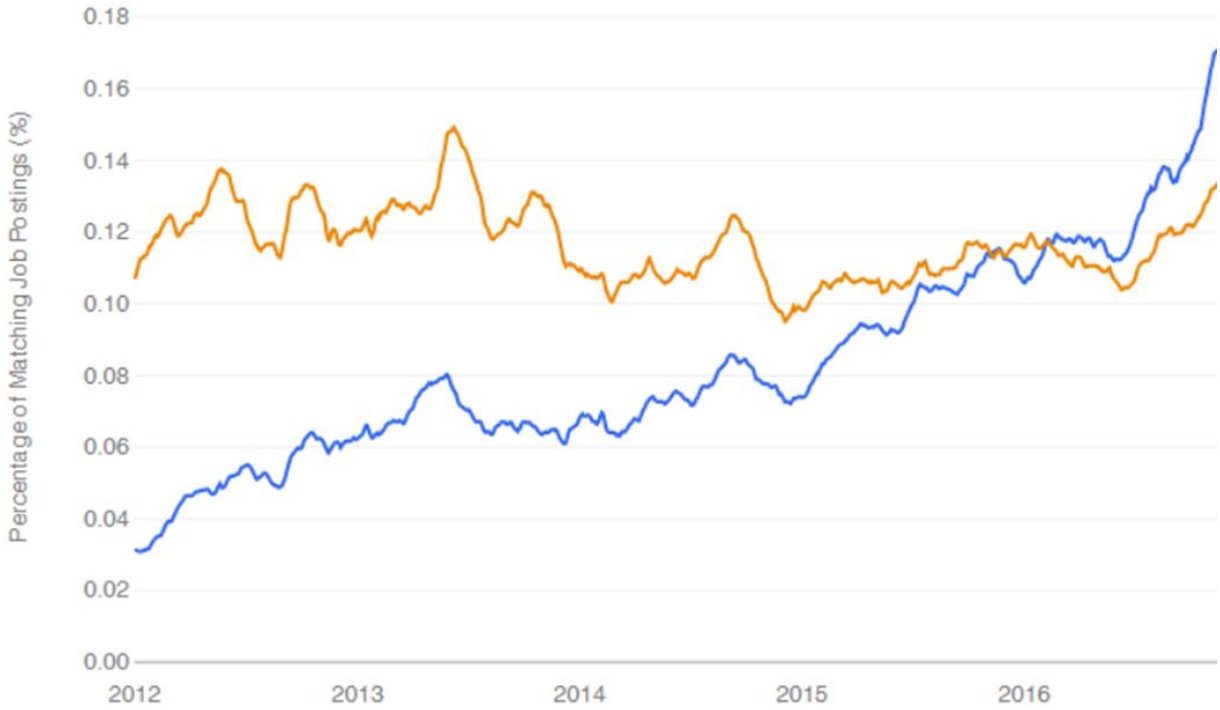

Fig 1. Percentage of Matching job postings (R in Blue, SAS in Orange).

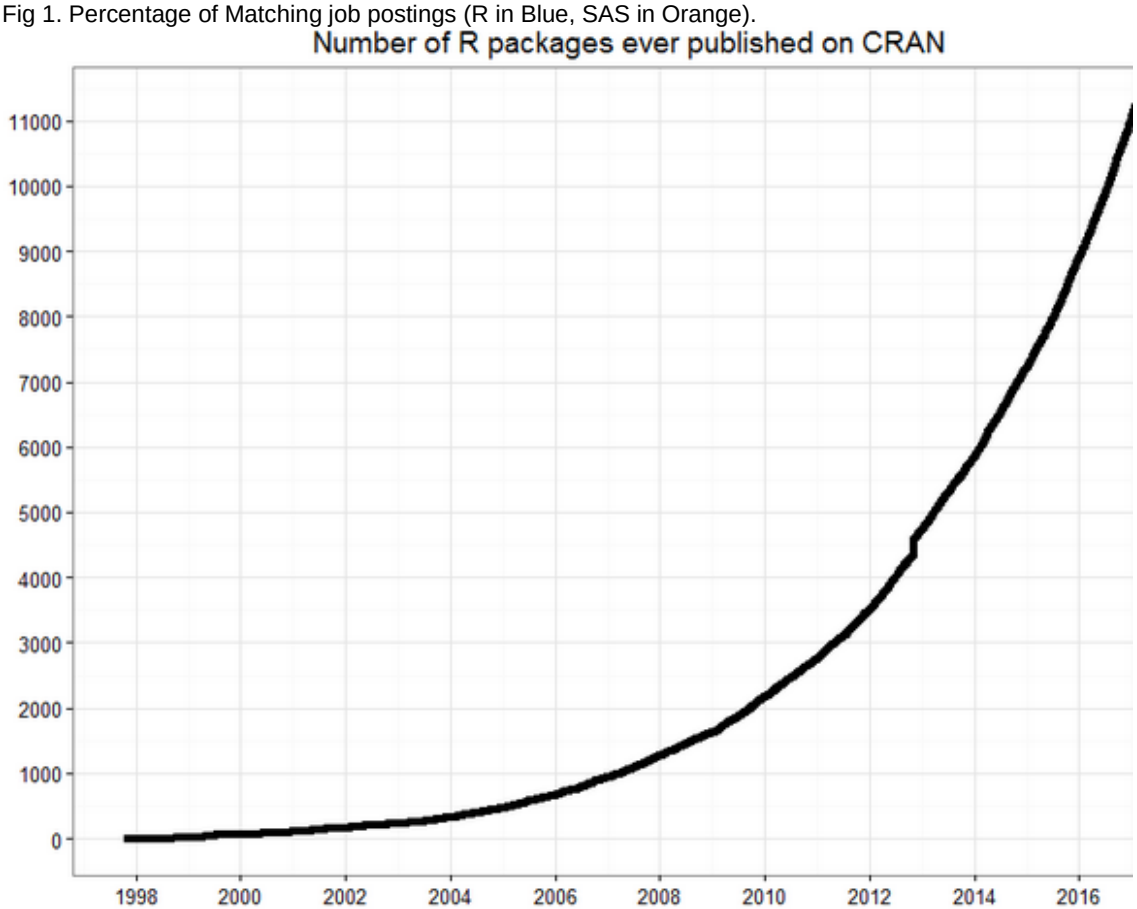

Fig 2. Number of R packages on CRAN

PhUSE US Connect 2018

DH03

R IN COMMERCIAL APPLICATIONS

Oracle has adopted R as a language and environment to support Statisticians, Data Analysts, and Data Scientists in performing statistical data analysis and advanced analytics, as well as generating sophisticated graphs. In addressing the enterprise and the need to analyse Big Data, Oracle provides R integration through five key technologies:

- Oracle R Distribution - Oracle's supported redistribution of open source R, provided as a free download from Oracle, enhanced with dynamic loading of high performance linear algebra libraries.
- ROracle - An open source R package, maintained by Oracle and enhanced to use the Oracle Call Interface (OCI) libraries to handle database connections - providing a high-performance, native C-language interface to Oracle Database.
- Oracle Analytics Cloud and the Data Visualization Desktop – Both use R for their Advanced Analytics and Machine Learning functions, allowing users to leverage existing R packages and upload their own to power their analyses. Visualizations created locally in the Data Visualization Desktop can be loaded to the Oracle Analytics Cloud, along with their R algorithms.
- Oracle R Enterprise - Integration of R with Oracle Database. Oracle R Enterprise makes the open source R statistical programming language and environment ready for the enterprise with scalability, performance, and ease of production deployment.
- Oracle R Advanced Analytics for Hadoop - High performance native access to the Hadoop Distributed File System (HDFS) and MapReduce programming framework for R users.

CERN (THE EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH)

Established in 1954, CERN (the European Organization for Nuclear Research) is the largest particle-physics laboratory in the world. Most famously, CERN is home to the Large Hadron Collider, the most powerful particle accelerator in existence. CERN uses big data, cloud computing, and analytics to help researchers unravel the mysteries of the universe, one petabyte at a time. Since 2003, CERN and Oracle have also partnered to drive innovation in ICT through CERN openlab.

The Large Hadron Collider is one of the most complex machines ever built. In addition to the petabytes of physics data it produces by smashing particles together at close to the speed of light, its control systems produce vast quantities of information. Analyzing these data streams and extracting key insights is vital in making sure researchers at the laboratory are able to continue pushing back the frontiers of our knowledge about the universe.

CERN is exploring ways the organization can efficiently and intelligently analyze the technical engineering information derived from around one million signals originating from its accelerator complex. The CERN team is building Machine Learning models to predict potential failures. These models use R (open-source distribution) and run in Oracle Database.

NHS BUSINESS SERVICES AUTHORITY

The NHS (UK National Health Service) is the largest and oldest single-payer healthcare system in the world. It provides healthcare to every legal resident in the United Kingdom, and free emergency treatment for everyone, including visitors.

The NHS Business Services Authority learned to make the most of its data thanks to Oracle learning labs and analytics tools, and identified huge potential savings. The organization was able to optimize treatment while reducing the use of less-effective medical procedures, and identify potential savings of US$156 million within three months. Savings of over US$1 billion have been identified to date.

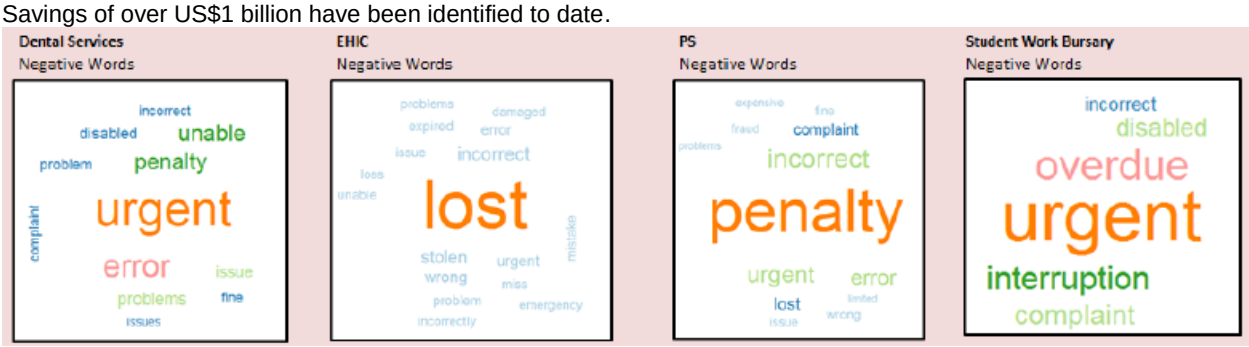

Figure 3: Feedback analysis using R

"*The overall solution is very fast, and our investment very quickly provided value. "The NHS sits on billions of data points that have the potential to deliver tremendous value to the wider healthcare system in the UK when combined*

and analyzed effectively. We can now do so much more with our data, resulting in significant savings for the NHS as a whole." We can now do so much more with our data, resulting in significant savings for the NHS as a whole."
Nina Monckton, NHS Business Services Authority

R IN CROSS STUDY ANALYSIS

Clinical Data Scientists can use pooled data in R from multiple study databases for their visualizations, and train and apply Machine Learning workflows over ever larger datasets. Oracle Health Sciences conducted a three-month proof of concept project to analyse SDTM data from two different studies, both conformed in DMW and available as secure Business Area data in the Oracle database. The team used the Oracle R Distribution 3.1.1 to prepare the analysis:

1. Connect to standardized SDTM data from multiple studies in DMW using secure Business Areas
2. Combine data across multiple studies using simple rbind procedure on each set of conformed domains
3. Create complex visualizations in R using publicly available packages
4. Train predictive analytics algorithms using publicly available packages
5. Apply term analysis using publicly available packages
6. Export to SAS V5 xpt using publicly available packages

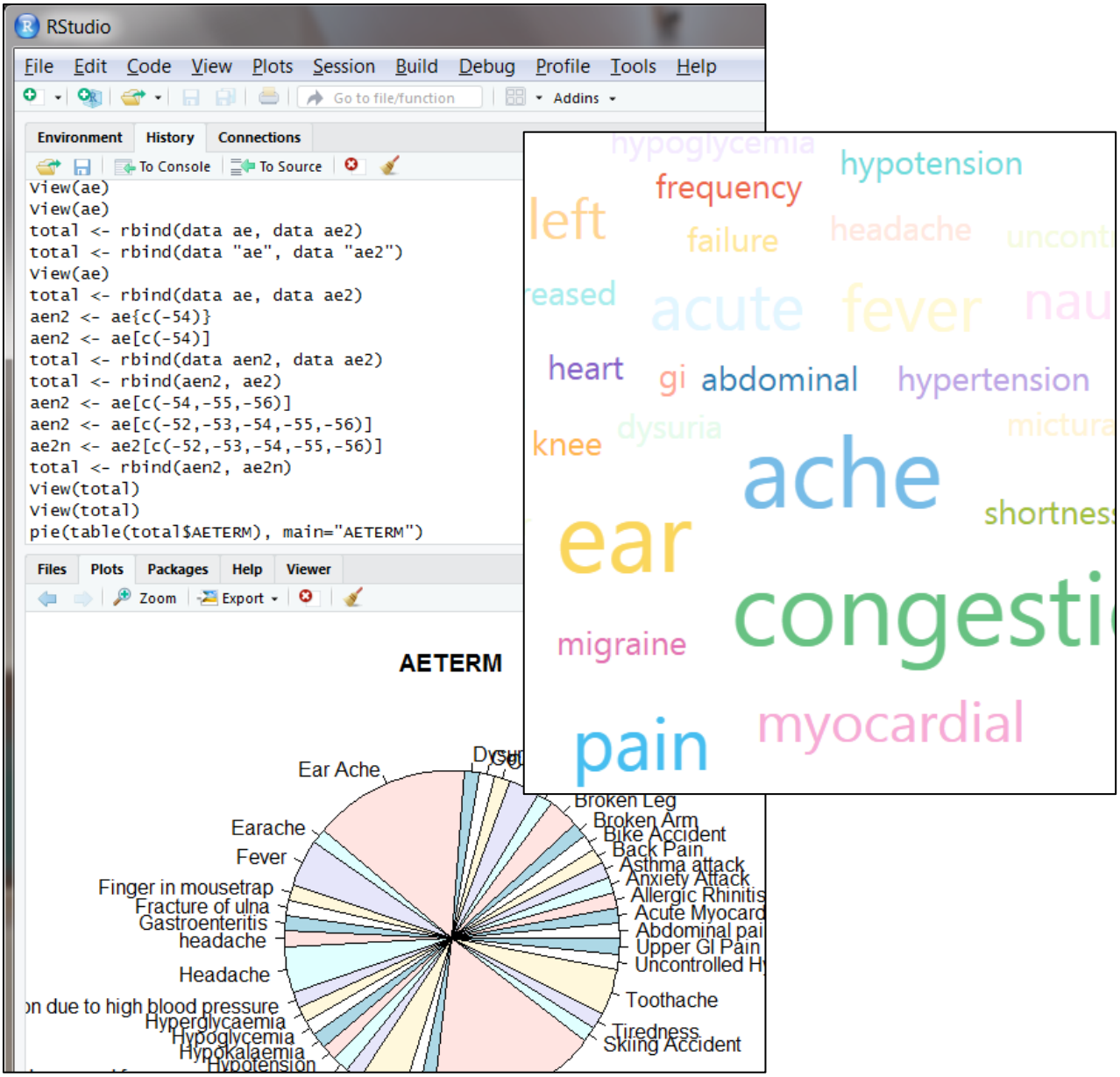

Fig 4. Cross Study AE Term Analysis Using R

Outcomes of Cross Study Analysis POC:

Positives	**Negatives**
Simple to connect to DMW data 1. Data blinding supported 2. Blind break recorded in source system	More complicated to combine data across multiple studies where structures are not identical
Easy to combine data across multiple studies where structures are the same	Performance was limited using remote database and local Rstudio 1. Rstudio server is also available 2. R is also available in the Oracle database
Possible to train predictive analytics algorithms	Adding bespoke packages can be time-consuming
Easy to create and share complex visualizations in R	Bespoke package had to be created to export the combined data into SAS
Possible to apply term analysis over combined sources	

In summary, R was an excellent tool to connect to the conformed DMW data, and allowed us to pool dataframes (datasets) using only a few keystrokes. Publicly available visualizations and algorithms further enhance the experience. It was a little time consuming to add a bespoke package of our own to allow export to SAS V5 xpt files, but there is a huge amount of support available, and as R is open source you can always to take advantage of code developed by others to speed your development.

HEALTHCARE ANALYSIS

R can support predictive analytics, such as hospital readmission rates:

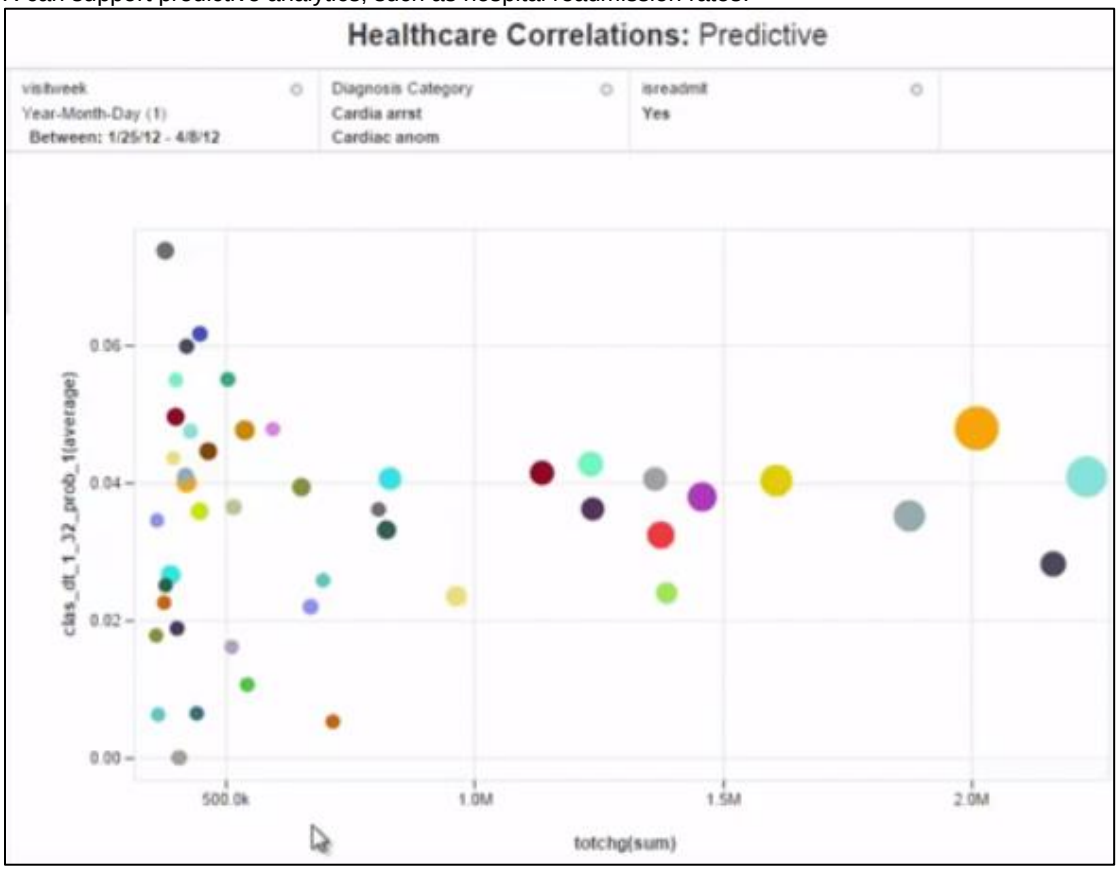

Fig 5. Hospital Readmission Rate Analysis Using R

R also offers the potential to leverage Machine Learning, such as heart disease predictions:

Model - Heart disease likelyhood Training - CART

General | Quality | Related

F1 Value Model Accuracy 47%

Actual Values \ Predicted Values	Absent	Less Likely	Likely	Others	Total
Absent	22	0	0	15	37 (62%)
Less Likely	4	0	0	4	8 (13%)
Likely	2	0	0	6	8 (13%)
Others	1	0	0	6	7 (12%)
Total	29 (48%)	0 (0%)	0 (0%)	31 (52%)	60 (100%)

Fig 6. Machine Learning Algorithm Used for Heart Disease Predictions – Confusion Matrix

REGULATORY CONSIDERATIONS

The FDA's Statistical Software Clarifying Statement declares that any suitable software can be used in a regulatory submission. Some data-exchange regulations do require the use of the XPT file format, which is an open standard, not restricted to SAS. XPT files can be read into R with the read.xport function, and data can be exported with the write.xport function in the SASxport package, while RStudio, a popular editor for R, uses the Haven package to import SAS datasets.

The R Foundation also provides guidance on how R complies with other FDA regulations: Regulatory Compliance and Validation Issues - A Guidance Document for the Use of R in Regulated Clinical Trial Environments.

R – THE FUTURE

- R use is clearly growing across many industries and it one of the key tools for today's Clinical Data Scientist.
- R is embedded in many leading industry solutions.
- R can power Machine Learning and Artificial Intelligence.
- The availability of a commercial distribution of R can re-assure users in even highly regulated industries.
- Confirmation from the FDA that it can be used to analyse clinical studies leaves no barriers to R adoption across the clinical trial lifecycle and beyond.

REFERENCES

Devices, Big Data and Real World Evidence, Collinson, PhUSE 2017
Clinical Data in Business Intelligence, Collinson, PhUSE 2016
SDTM in Business Intelligence, Collinson, PhUSE 2014
https://www.youtube.com/watch?v=6YR_YPp70cU
https://www.youtube.com/watch?v=IichF5pBt_U
https://www.r-project.org/doc/R-FDA.pdf

ACKNOWLEDGMENTS

https://www.r-project.org/foundation/

RECOMMENDED READING

https://www.oracle.com/solutions/business-analytics/data-visualization/library.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:
Oracle Health Sciences
100 Crosby Drive
Bedford,
MA 01730, USA