R Programming in clinical trial data analysis

23 July 2020 | Features | By Shrishaila Patil

It is vital that we prepare our workforces with the necessary skillsets to meet future needs

In the last few years, Data Science has been fuelling powerful business decisions taken by Industry leaders. Data scientists are story tellers. They often need to dig into Data, clean, transform, build and validate models, understand pattern, generate insights and, most importantly, communicate results effectively.

In the field of Statistics, Analytics and Visualization, in addition to SAS (**Statistical Analysis System)**, most talked about languages are R and Python. This article highlights the current status, the observed challenges of R, proposed approaches for the risk assessment of R packages, mitigation, and implementation for Clinical Trial Data Analysis.

So, what is the Need of the Hour? It is of paramount importance that we understand the bigger picture for the Life Sciences Industry.

"**The Life Sciences Industry needs greater innovation in order to reach patients faster, with affordable drug prices and improved accessibility. With the ongoing COVID-19 crisis, this is now more important than ever before".**

The industry is seeking better alternative Technologies and Tools, which are sustainable and can provide optimal solutions while effectively addressing Industry challenges. There are a number of questions to be addressed:

- Are our current Tools outdated?
- Do we have an alternate solution for SAS, in order to avoid high license and maintenance cost?
- Is the Industry ready for R, Python, or another tool?

What is certain, is that we need efficient Data Science Technologies and Tools that can help us to manage Data Lake (Example: Big Data and Real-World Data) while being able to process it faster and with accuracy. Efficiency in Data Analysis results in greater Insights about data and can help improve decision making across Drug Development.

Innovation is needed in order to move away from any traditional inefficient processes and tools, towards efficient, simple, easy to implement, reliable and cost-effective solutions. Collaboration across Industry stakeholders is needed to develop better technology ecosystems and to agree on Validation and Regulatory benchmarks.

It is vital that we prepare our workforces with the necessary skillsets to meet future needs.

Current Trends of R in Pharma and CROs:

R is a language and environment for statistical computing and graphics. It is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. As an open source software, R gets huge support from the Data Community. Source code availability provides superior and thorough documentation. There are approximately 2 Million users worldwide for R.

In order to understand how R is used across Pharma and CRO Industry in India, a survey was conducted on social media. More than 25 Companies and 250+ participants engaged in this survey. Looking at the current industry trends, R usage is less than 10% in activities related to Pharma Regulatory Submissions. However, R is extensively used in public health projects, healthcare economics, exploratory/scientific analysis, trend identification, generation of Plots/Graphs, specific Statistical analysis and machine learning. R is not widely used for creation of CDISC (SDTM, ADaM) datasets.

One of the most common questions from the Programming community is "Should we replace SAS with R, or use both, or another language (Python)?". There is no one size fits all, instead of choosing between SAS or R or Python, we should leverage the best from each of these programming languages to solve appropriate Data science problems.

We have few Early Adopters of R, and they have experienced some challenges. Ensuring regulatory compliance of R packages is one of the most common. If R has to be considered for doing work related to regulatory submissions, one needs to do Risk assessment of R packages, feasibility analysis, and establish process for R usage through Pilot projects with the necessary documentation.

Often, we end up using different technologies and it can become difficult to integrate when needed.

The Technology space continues to expand, and it is paramount that we stay ahead, in terms of learning curve, in order to take advantage of the cutting-edge solutions that are required for Data science problem solving.

In India, workforce retraining is one of the major strategies that Companies must embark on for the next 2 to 3 years to ensure their employees are upskilled with the latest technology and tools.

R Validation Hub: Enabling Use of R in Regulatory Setting:

R is free, but it is an investment. The main challenge of using R is ensuring validation documentation. R needs to be programmed.

In May 2015, the US FDA released a Statistical Software Clarifying Statement. The FDA does not require use of any specific software for statistical analysis. However, software packages used for statistical analysis should be fully documented in the submission, including version and build identification. In addition, documentation of appropriate software testing procedures should be readily available.

In March 2018, the FDA released the Study Data Technical Conformance Guide. Delivering Software Programs, paragraph 4.1.2.10 states: "**Sponsors should provide the software programs used to create all ADaM datasets and generate tables and figures associated with primary and secondary efficacy analyses.** Furthermore, sponsors should submit software programs used to generate additional information included in Section 14 CLINICAL STUDIES of the Prescribing Information (PI)26 if applicable. The specific software utilized should be specified in the ADRG (Analysis Data Reviewer's Guide).

The main purpose of requesting the submission of these programs is to understand the process by which the variables for the respective analyses were created and to confirm the analysis algorithms. Sponsors should submit software programs in ASCII text format; however, executable file extensions should not be used.

R Validation hub is a Cross Industry Initiative. The mission is to enable the use of R by the Bio-Pharmaceutical Industry in a regulatory setting, where the output may be used in submissions to regulatory agencies.

The R Validation Hub comprises of participants from across the pharmaceutical industry (Abbvie, Amgen, Astellas, Bayer, Boehringer-Ingelheim, Celgene, Eli Lilly, FDA, Genentech, Gilead, GSK, Johnson & Johnson, Merck, Novartis, Novo Nordisk, Pfizer, Roche, RStudio, Sanofi, Teva Pharmaceutical Industries Ltd, and many more). Participants contribute to the effort through regular group meetings, as well as the various workstreams that make up the project.

The focus of this group is on designing a framework that assesses the quality of an R package (Contributed by volunteers) and to create a repository of "accepted" packages.

Shrishaila Patil, Vice President, Statistical Programming, Navitas Data Sciences, a part of Navitas Life Sciences (a TAKE Solutions Enterprise), Bengaluru