PharmaSUG 2022 - Paper SI-055

# Implementing and Achieving Organizational R Training Objectives

Jeff Cheng, Abhilash Chimbirithy, Amy Gillespie, and Yilong Zhang
Merck & Co., Inc., Kenilworth, NJ, USA

## ABSTRACT

Clinical data analysis and reporting for a clinical study report, submission, and internal decision-making are complex activities. The work requires large amounts of data to be processed, analyzed, and reported according to regulatory and departmental processes. As such, tools which help maintain process control, increase productivity, and sustain resourcing are in demand by statistical programming organizations. With many open-source libraries for data manipulation, statistical calculation, and visualization functions, R Statistical Programming Language is a tool that has gained popularity within the pharmaceutical industry. In this paper, we will highlight how Merck implemented its training strategy for R to achieve essential R proficiency in 75% of its statistical programming staff within one calendar year. We will share how the R curriculum was selected, planned, and implemented, which resulted in a robust training strategy that successfully achieved the training objective. We will also discuss the challenges and benefits of this R training strategy which could be helpful for other organizations that might be contemplating a similar training initiative for their employees.

## INTRODUCTION

R is a popular open-source programming language for data manipulation, visualization, and statistical analysis. There is a growing trend of using R for clinical development in the pharmaceutical industry. Using R to improve analytical and administrative efficiency is particularly attractive.

Biostatistics and Research Decision Sciences (BARDS) is a highly distinguished department within Merck's renowned Research & Development Division. BARDS members partner with other disciplines within Merck to provide statistical analysis, reporting, and data interpretation to enable informed decision-making in drug and vaccine development. Within BARDS, approximately 300 statistical programmers located in the US, Europe, and Asia play a critical role in preparing and delivering statistical analysis results. One of BARDS departmental goals in 2021 was to upskill at least 75% of its statistical programmers in R programming. This goal initiated an important journey towards becoming a multi-lingual statistical programming organization.

Our R training approach was multidimensional; however our primary training mode was through internally presented training conducted by fellow colleagues. The training introduced data manipulation skills, focusing on utilizing functions from the tidyverse R package, hence equipping the trainees with working knowledge of using R for data analysis and visualization. The target audience was experienced SAS programmers who were familiar with clinical development.

Most online R training courses are designed for general audiences. Whereas an in-house designed training can be tailored using data and scenarios more relevant to the trainees' day to day work and responsibilities. This relevancy helps the trainees reduce their learning curve with familiar problem scenarios that they can easily relate to. It will also increase motivation to learn the course materials, inspire innovation, and create business value by adopting new functionalities in R.

Another important goal of the internal R training program is to build a user community by forming small cohorts conducting sessions similar to book clubs. Within the cohort, an experienced R developer from the department acts as the coach, guiding four to six trainees through the training. In addition, the effort was

made to group a coach with trainees from a similar functional area to encourage better communication and discussions within the cohort.

With the philosophy of effective learning in mind, we actively involved the trainees by assigning pre-work for each training session and by having the trainees present the course material to their peers during a training session. With a complete set of training materials provided, each trainee had the opportunity to lead the presentation and discussion of one R topic under the guidance of the training coach within each cohort. This "learning-by-doing" strategy also helps to develop skills and confidence in researching, experimenting, collaborating, and presenting to peers while learning R.

## TRAINING DEVELOPMENT

### CURRICULUM

We leveraged many publicly available training materials to reduce the effort in preparing our training curriculum. The slides presented in the cohort were adapted from Dr. Garrett Grolemund's Master the tidyverse workshop. These slides provide comprehensive introduction to the tidyverse packages and concepts of data manipulation with detailed examples. We updated slides based on examples using clinical data to make the material more relevant to the trainees' job role. The famous R for Data Science [1] book was also selected as the textbook for the training. Lastly, a few RStudio webinars were recommended for trainees to gain additional knowledge for the R topics covered.

Adoption of publicly available training resources reduced a significant amount of time and effort required to prepare materials for our in-house training, allowing us to focus our efforts on the following items:

- Prepare a training schedule with reading and exercise materials before, during, and after a training session.
- Select relevant training topics suitable for 5-6 one-hour training sessions.
- Develop easy-to-follow examples based on datasets typically used for clinical trial analysis & reporting (CDISC ADaM & SDTM).
- Develop a milestone project for data manipulation and visualization.

In the Appendix, a sample training schedule is provided with recommendations on training materials to be reviewed before and after each session. The original training slides in the Master the tidyverse workshop contain ten different topics, which were condensed by picking five essential topics to meet our training goals:

1. Data transformation with dplyr
2. Tabular data with tidyr
3. R data types
4. Data iteration and vectorization
5. Visualization with ggplot2

With this condensed version, our target was to complete the training within six to eight group sessions of one hour each. We provided supplemental material for pre and post cohort sessions to enhance the trainee’s knowledge of data manipulation using tidyverse. We made the training material more relevant to a trainee's job function by developing example code using CDISC ADaM datasets. For instance, while illustrating the dplyr::filter() function to subset observations, a simple example of filtering subjects with age greater than 80 years old in an ADSL dataset is used:

```
filter(adsl, AGE > 80)
```

These work-related yet straightforward examples help our trainees to quickly digest the information and connect the knowledge with their day-to-day activities. Training material that is relevant to a trainee's day-to-day activities and responsibilities help to reduce the learning curve and improve motivation for learning. At the end of each session, additional background material is provided for motivated trainees. For example, we recommend that motivated trainees read both internal and external examples in creating an adverse event table in RTF format in one of the sessions. (e.g., https://r4csr.org/specific-ae.html)

At the end of all training sessions, a milestone project with various data exploration scenarios was also developed to help the trainees solve real-world problems by applying their learning. One such scenario is for a trainee to find out how many subjects had serious AEs by treatment group by exploring an ADAE dataset.

## TRAINING PLAN

On average, it takes three months for a cohort of four to six trainees to complete all five to eight training sessions. The time commitment for each session is approximately 3-4 hours per trainee, which amounts to 18 – 28 hours in total for the entire training.

A different trainee will review and present existing slides of one of the R topics to the cohort in each training session. The coach may determine the frequency of sessions (e.g., one session per week) based on the availability of the cohort members. It is essential for the trainee to review the materials before the presentation so the cohort can maximize the time for hands-on exercises during the session.

The main objective of the training is to help the trainees obtain a working knowledge of R programming i.e., syntax, library functions, programming environment, and learning resources. Our training is not a comprehensive R programming course, and our goal is only to provide sufficient exposure to motivate SAS programmers to continue their learning with strong support from the R community within our department. After the initial training, many trainees become key contributors and maintainers for our internal standard R packages.

We also used the training program as a platform to promote the use of version control tools. All training material is developed and maintained by an internal training team using Bitbucket. As part of the RStudio account setup, trainees are asked to clone the course materials as a project from Bitbucket. While this is not a full exposure to Git for version control, the strategy helps us spread agile project management concepts. Some study teams within Merck are starting to adopt this practice by managing their clinical project using Jira and Confluence [2].

Even though the coach-led training format has many benefits, it is somewhat time-intensive and requires scheduled commitment. Hence for trainees that are time-constrained, a self-paced online Coursera alternative offered by Johns Hopkin University titled: "Data Science Specialization", is provided to trainees who need time flexibility in their R learnings.

In general, we recommend trainees register for the in-house R training as it is more aligned to our internal processes and standards. The internal training also provides insights to the R platform at Merck, helping trainees to move quickly towards supporting R-related projects and initiatives within the company.

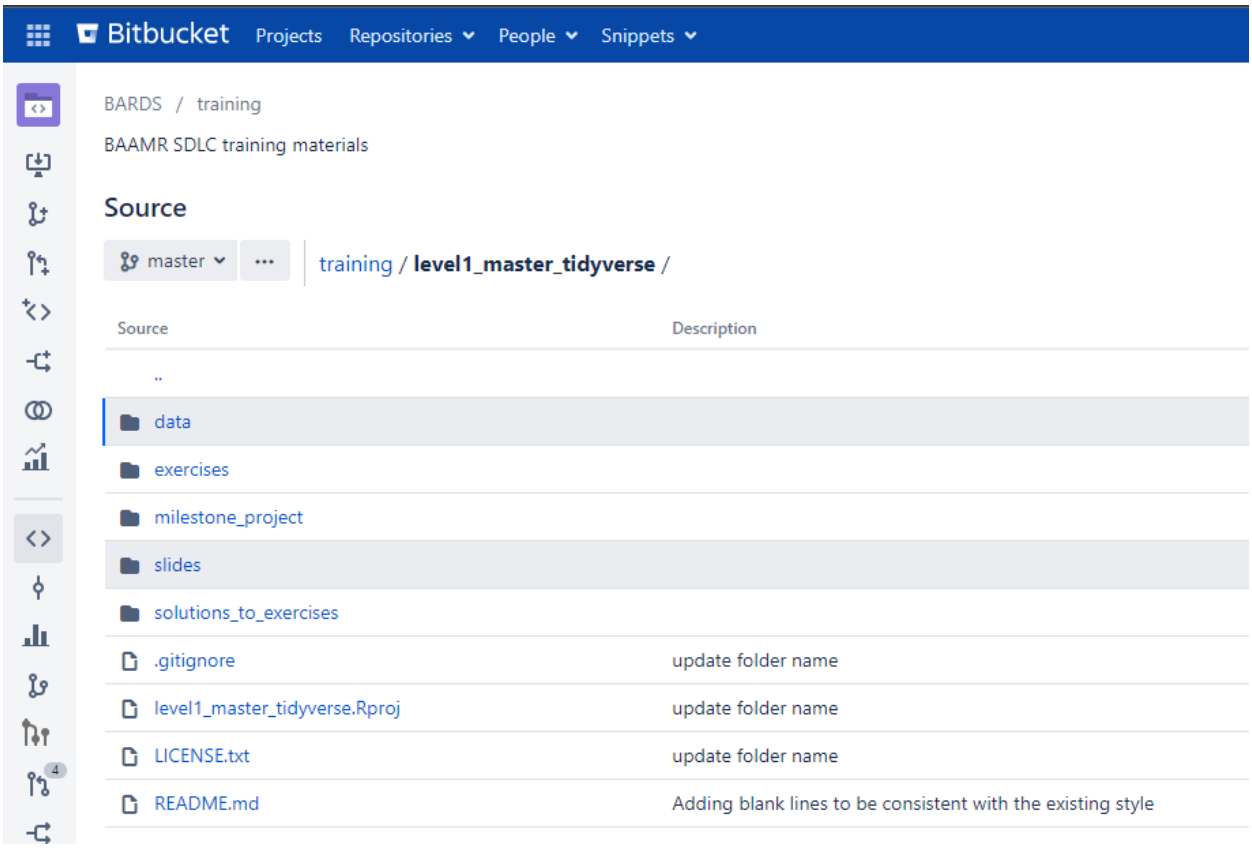


**Figure 1. Bitbucket Example: R Level 1 training course website**

## COACH

The training is designed to minimize the workload of a coach because coaches volunteered to support the training program while still having to maintain their project work as a priority. For an experienced coach, we expect them to commit one hour to conduct the session and 30-minutes to manage the cohort per session. In addition, previous learning session recordings, email templates, and kick-off meeting slides are provided to coaches to assist them in facilitating the sessions and reduce the workload required for this training.

Each coach has full control over the cohort's flexibility, training schedule, and training material. For example, a coach can split a topic into multiple sessions if the material cannot be covered in a one-hour session. Coaches are expected to encourage more discussion and support amongst the trainees and to fill in gaps during a learning session. The coach also helps trainees identify bugs during exercises and review milestone projects. Weekly 30-minute office hours are also scheduled to answer trainees' questions. Coaches take turns conducting the weekly office hours for a 3-month period.

For coaches-in-training, we assign them the role of "coach-assistant" within an ongoing cohort, so they can observe how an experienced coach manages a cohort while preparing to be independent coaches in the future. We recommend all new coaches watch the RStudio 2019 conference keynote talk on explicit direct instruction in programming education by Prof. Hermans prior to joining the training.

## TRAINEE

Based on the design of a 3-month training period, trainee recruitment occurs on a quarterly basis. At the beginning of 2021, we collected training preferences from managers of statistical programming groups by providing our training plan and training goals. In response, the managers specified the R training preferences and needs of their direct reports. Using this input, we began assigning trainees to a specific quarter. After obtaining a good estimate of the number of trainees in each quarter, a call for volunteer coaches was sent out to a list of experienced R developers. Trainees and coaches were then matched into cohorts of four to six members, each based on their functional area and geographical location.

The training is designed to maximize the hands-on experience of a trainee. We strongly encourage each trainee to complete the reading materials before each session and actively complete the exercises during the session. The trainee presenter should guide their peer trainees through ~90% of the materials in the slides.

## LOGISTICS

A core team is formed to manage the training program. The responsibilities include overseeing the development of the course materials and training plan, recruiting trainees and coaches, monitoring the overall training status, and communicating the findings to the leadership team.

Each quarter, the following steps are taken to set the training in motion:
Notification emails are sent to the assigned trainees to:

- Confirm availability
- Provide time commitment estimates and training expectations
- Provide links to the training material repository
- Provide information on how to obtain an RStudio account

Trainees must obtain RStudio Server accounts prior to starting the training. We provide access to our internal cloud-based RStudio server; alternatively, trainees can use RStudio Cloud (https://rstudio.cloud/). Following the instructions, trainees will complete new user setup and download training materials from the Bitbucket training repository using Git.

The coach initiates quarterly kick-off meetings for each cohort to go over administrative details and align training expectations and responsibilities with the trainee at the beginning of the training. Weekly office hours are conducted rotationally by the coaches to answer trainees' questions. Coaches are asked to report any unexpected challenges during the training to ensure on-time completion.

After the training is completed, the training completion records are documented in the company's learning management system for future reference. Post-training completion surveys are conducted to provide feedback on the topics presented during training. We strive to continuously identify learning gaps to improve the training experience, process, and materials based on the feedback.

## CONCLUSION

The tools available for clinical data analysis and reporting are constantly evolving. Hence the ability to assess, learn and adopt new productivity-improving tools is essential for a statistical organization to continuously perform at a high level.

By developing time-flexible, content-relevant, and resource-efficient introductory R training curricula, we hope to better take advantage of the many capabilities of R in data processing, analysis, and visualization,

as realized by its vast array of function libraries. In addition, the experience of researching, experimenting, and presenting new information through group learning will also help individuals efficiently adopt new tools in the future.

With strong support from both the management team and motivated members of the BARDS organization, we successfully guided 41 cohorts with 215 trainees through R Level 1 training in 2021. This exceeded the overall department goal in 2021 and we continued to have strong enrollment in 2022.

In anticipation of a broader adaptation of R in regulatory filings and to make future R project development more robust, efficient, and user friendly, we also provided training on R package development in a similar structure. In addition, we are actively developing training materials for our internal analysis and reporting process, visualization, and shiny development.

In the early stages of formalizing our R adoption strategy, we did experience some resistance from SAS programmers to register for R training. Most resistance was primarily due to change, uncertainty and hesitation about the organization's multi-year and multi-lingual programming strategy. Many questions were received by the trainees comparing the pros and cons of SAS and R and the need for change. While change and evolution at this scale can be challenging, it is important to recognize the importance of continuous learning and skillset expansion.

In summary, we hope our R curriculum development experiences, with the objective of giving the trainees exposure in using R and its development tools with different training options, will help those thinking of incorporating R in their own clinical data processing workflow by upskilling their workforce.

## REFERENCES


[1] Garrett Grolemund, Hadley Wickham (2017). R for Data Science. O'Reilly Media.

[2] Sarad Nepal, Uday Preetham Palukuru, Peikun Wu, Madhusudhan Ginnaram, Ruchitbhai Patel, Abhilash Vasu Chimbirithy, Changhong Shi, Yilong Zhang. 2021. Agile Project Management in Analysis and Reporting of Late Stage Clinical Trials. PharmaSUG 2021 - Paper SI-083


## ACKNOWLEDGMENTS


The authors would like to thank the management teams from Merck & Co., Inc., Kenilworth, NJ, USA, for their advice on this paper/presentation.


## CONTACT INFORMATION

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


Jeff Cheng
Merck & Co., Inc., Kenilworth, NJ, USA
email: jeff.cheng@merck.com

Abhilash Chimbirithy
Merck & Co., Inc., Kenilworth, NJ, USA
email: chimbirithy_abhilash@merck.com

Amy Gillespie
Merck & Co., Inc., Kenilworth, NJ, USA

email: amy_gillespie@merck.com

Yilong Zhang, Ph.D.
Merck & Co., Inc., Kenilworth, NJ, USA
email: yilong.zhang@merck.com

## APPENDIX:

## SAMPLE TRAINING SCHEDULE

### Level 1: Master the tidyverse

***Introduction***

Welcome to the level 1 training for **Master the Tidyverse**. The training materials were adopted from Dr. Garrett Grolemund's master the tidyverse workshop

***Information for New Coach***

- Explicit Direct Instruction in Programming Education - Felienne
- Coach is responsible for setting up meetings.
- If a presenter is unavailable for their assigned week, coach will work with the presenter to find a replacement.
- Initiate a training kick-off meeting to define order of presenter.

***Information for Trainee***

To enhance the training experience, please complete items below before the first session.

1. Be able to clone/download the training project and open in RStudio.

### Session 1 Transform Data: Dplyr

Before the session:

- Watch Data wrangling with R and RStudio (~ 1 hour)
- Watch Getting Data into R (~ 1 hour)
- Read Cheatsheat for SAS Users
- Read Chapter 5 of R for Data Science
- Read Chapter 13 of R for Data Science

Self-Learning Material

- Transforming Data with dplyr

During the session:

Present the slides in slides/01-Transform-Data.pptx

### Session 2 Tidy Data:

Before the session:

- Watch Getting started with R Markdown (1 hour)
- Read Chapter 12 of R for Data Science
- Read Chapters 14, 15, and 16 of R for Data Science

Self-Learning Material

- Tidy Data with tidyr

During the session:

Present the slides in slides/02-Tidy-Data.pptx

Additional A&R Examples (Optional)

- Create an adverse events table following https://r4csr.org/specific-ae.html

## Session 3 Data Manipulation

Before the session:

- Watch The Grammar and Graphics of Data Science
- Read Chapter 1 and 4 of R for Data Science
- Optional Reading: Chapters 14 (Strings), 15(Factors), and 16 (Dates and Times)

During one hour session:

Present the slides in slides/03-Data-Types.pptx

Additional A&R Examples (Optional)

- Create a baseline table following https://r4csr.org/baseline-characteristics.html

## Session 4 Tidy Data: Iteration and purrr

Before the session:

- Watch A Gentle Introduction to Tidy Statistics in R
- Read Chapter 21 of R for Data Science

During the session:

Present the slides in slides/04-Iteration.pptx

Additional A&R Examples (Optional)

- Create an efficacy table following https://r4csr.org/efficacy-table.html \

## Session 5 Data Visualization with RStudio and ggplot2

Before the session:

- Watch Rstudio General Data Science Overview (44 min)
- Watch Rstudio The Grammar and Graphics of Data Science (58 min)
- Read Chapter 3 of R for Data Science

Additional A&R Examples (Optional)

- Create a figure following https://r4csr.org/efficacy-figure.html

During one hour session:

- Present the slides in slides/05-Visualize-Data.pptx