PharmaSUG 2025 - Paper OS-195

Building a Scalable Training Platform for R: Empowering Analytics Excellence in Corporate Initiatives

Michelle Page-Lopez and Martyn Walker, Syneos Health

Abstract

Building a scalable training platform for R within a Contract Research Organization (CRO) addresses the growing demand for advanced analytical tools in clinical research. As the industry shifts toward open-source programming, equipping staff with the necessary R skills becomes critical. This paper highlights the creation of a flexible, accessible, and comprehensive training framework that fosters skill development while meeting organizational needs. The training platform leverages modern e-learning technologies, combining self-paced modules, interactive coding exercises, and real-world case studies. We showcase the *{learnr}* package in R, which enables the creation of hands-on training content by combining narrative text, R code, quizzes, and visualizations. This approach is ideal for creating tutorials that are customizable, scalable, and accessible through cloud-based infrastructure. Tailored to diverse proficiency levels, the curriculum covers foundational R programming, Study Data Tabulation Models (SDTM), Analysis Data Models (ADaMs), and Tables, Figures, and Listings (TFLs). Finally, we examine the challenges associated with developing an in-house training program from the ground up and explore potential future applications of such a platform. Training staff in R is vital for clinical research, equipping teams with the skills to leverage its powerful analytical capabilities, meet organizational needs, and ensure data-driven decision-making.

Introduction

Building a training program from scratch can be daunting, especially in an industry where new and advanced analytics tools emerge faster than they can be adopted. Open-source coding for clinical trials is becoming the new standard as companies seek to enhance analytics while managing costs. Among these tools, R is increasingly accepted and preferred for driving studies toward regulatory submission.

Contract research organizations (CROs) must adapt their programming teams to meet this growing demand. While R has been around for decades, SAS has long dominated the statistical programming industry. As a result, most statistical programmers are fluent in SAS, but fewer have a comprehensive understanding of R and its full capabilities.

This gap underscores the need for an internal training program. To keep pace with industry demands for more complex statistical methods while optimizing costs, a corporate initiative was launched to develop a scalable training program for biostatisticians and statistical programmers.

This paper serves as a guide to creating, maintaining, and optimizing a scalable training platform, helping statistical professionals of all proficiencies master R and stay ahead in the evolving landscape of clinical trial analytics.

Creating a Training Platform

Leveraging the {*learnr}* Package

The *{learnr}* package in RStudio offers a powerful way to transform static R Markdown documents into interactive, adaptive learning modules, an approach that significantly enhances both engagement and knowledge retention. By incorporating elements such as code exercises, multiple-choice questions, and even Shiny applications, *{learnr}* creates a dynamic and immersive learning experience that goes far beyond traditional training formats.

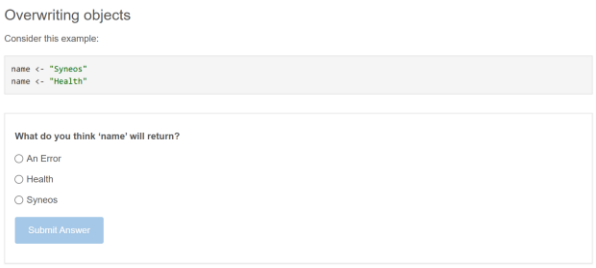
In our internal R training program, we leveraged *{learnr}* to create hands-on, scenario-driven exercises that reinforced key programming concepts. Rather than passively reading through code examples, users actively engaged with the material by writing and testing their own code, receiving immediate feedback, and revisiting concepts as needed. This interactive structure allowed for more self-directed learning, enabling users to work at their own pace while ensuring they achieved a solid understanding of foundational R concepts.

We primarily used code-based exercises and quiz-style multiple-choice questions to support learning throughout the training modules. Code exercises allowed users to apply programming techniques in real time, while quizzes served to check comprehension, reinforce best practices, and provide feedback.

The *{learnr}* package is freely available through the Comprehensive R Archive Network (CRAN), the central repository for R packages. To implement it, developers simply install the package within RStudio and load it into any R Markdown document using a standard *library(learnr)* call. Each tutorial is then authored in R Markdown using built-in *{learnr}* functions such as *question()* and *exercise(),* allowing for highly customizable training content.

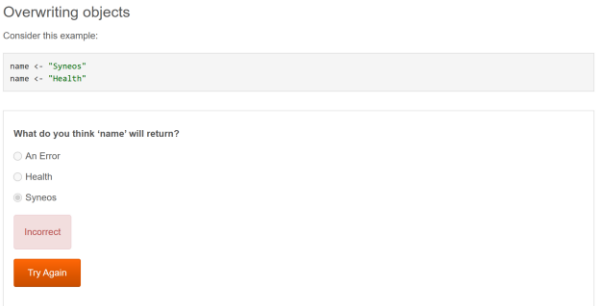
This approach proved especially effective in a corporate setting, where users have varying levels of experience and limited time. By moving away from static documents and slide decks and toward interactive, modular content, we created a more engaging, personalized training experience that encouraged deeper comprehension and long-term retention.

Below is an example of how *{learnr}* was implemented in one of our internal modules, demonstrating its ability to support both beginner-level instruction and advanced programming practice.



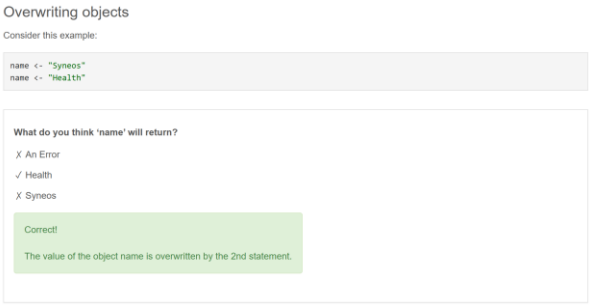
Display 1. Training Module Example for *{learnr}* Package Implementation

This question stems from the fundamental concept of overwriting objects in R. By leveraging *{learnr}'s* interactive multiple-choice functionality, we can effectively assess a trainee's understanding of the topic. If an incorrect answer is selected, the user receives immediate feedback and is prompted to try again, reinforcing learning through active engagement.



Display 2. Training Module Incorrect Answer Example for *{learnr}* Package Implementation

Once the correct answer is provided by the user, the below is shown:



Display 3. Training Module Correct Answer Example for *{learnr}* Package Implementation

The text in the green box is customizable within the *{learnr}* package. Sample code of this is provided:

```
1.  ### Overwriting objects
2.
3.  Consider this example:
4.
5.  ```{r}
6.  name <- "Syneos"
7.  name <- "Health"
8.  ```
9.
10. ```{r Q3, echo=FALSE}
11.
12. question("What do you think 'name' will return?",
13. answer("Syneos"),
14. answer("Health", correct = TRUE, message = "The value of the object name is
    overwritten by the 2nd statement."),
15. answer("An Error"),
16. allow_retry = TRUE,
17. random_answer_order = TRUE)
18. ```
```

Program 1. Sample *{learnr}* R code

In *{learnr}*, a multiple-choice question is created using the *question()* function, as shown in line 12 of the sample code. Possible answers are added with *answer()*, and the correct response is marked with *correct = TRUE*. Developers can also provide explanatory feedback using *message = ""* within *answer()*, as seen in line 14. In this example, "Health" is correct because, in R, an object's value is always determined by its last assignment, even if it was previously set to another value.

This approach was applied across all training topics, which will be explored further. The *{learnr}* package fosters engagement for learners of all skill levels.

Centralized Training Access

To ensure ease of access, usability, and a secure learning environment, the decision was made to host all R training materials on a dedicated SharePoint site designed specifically for the R training initiative. SharePoint was already well-established within the organization as a comprehensive knowledge management system, making it a natural fit for consolidating and distributing training resources.

This centralized hub serves as the entry point for all users participating in the program. It houses structured training pathways, links to interactive modules, downloadable exercises, and key documentation related to the program. Each training phase, R Basics, SDTM/ADaM, and TFLs, is clearly organized, allowing users to navigate content easily based on their current learning stage.

Additionally, SharePoint's built-in security and access controls ensure that only authorized users can access training content, which is particularly important when working with proprietary or internal-facing material. Version control features and user permissions help maintain the integrity of the training program as it evolves.

All interactive tutorials and learning modules, developed using the *{learnr}* package and hosted through the R publishing platform, are seamlessly linked within SharePoint. This integration allows users to launch modules directly from the portal, providing a streamlined and consistent experience regardless of their geographic location or technical background.

By leveraging SharePoint as the centralized platform, the training initiative benefits from both organizational alignment and scalability, enabling smooth onboarding, resource management, and cross-functional visibility into the program's offerings.

Curriculum

Since SAS has long been the dominant programming language in the clinical research industry, many experienced statistical programmers have built their careers around SAS-based workflows and have had limited exposure to R. In contrast, R has gained widespread popularity in academic settings due to its open-source nature, flexibility, and cost-effectiveness, making it a standard part of the curriculum in many post-secondary education programs. As a result, there is a noticeable divergence in skill sets: while most statistical programmers are fluent in SAS, many biostatisticians enter the industry with a solid foundation in R but less familiarity with SAS.

This creates a unique challenge when designing a training program intended for a diverse audience. Users may come in with varying degrees of comfort and proficiency - some with extensive programming experience but little exposure to R, and others with a working knowledge of R but limited experience in applying it within a clinical trial context.

To address this disparity, the training program was intentionally designed to be inclusive, scalable, and adaptable to different skill levels. To meet users where they are and ensure a gradual, logical progression of concepts, the curriculum is organized into three core phases:

1. R Basics - Introduces the fundamentals of R and its programming environment, serving as a foundation for new users or a refresher for those with limited experience.
2. SDTM and ADaM Development - Focuses on applying R in the context of clinical data standards, providing practical, hands-on examples for dataset creation and manipulation.
3. TFL Generation - Covers more advanced use cases, including the creation of Tables, Figures, and Listings, and techniques for outputting results in industry-ready formats.

This structured approach ensures that users can build on their knowledge progressively, regardless of their starting point. By tailoring the program to a wide range of backgrounds, the initiative aims to promote organization-wide fluency in R and empower teams to apply it effectively across the clinical trial lifecycle.

Phase 1: R Basics

The first category, **"**R Basics", introduces users to the RStudio interface and the concept of packages in R. From there, the training progresses to interactive lessons covering key fundamentals such as assigning values, working with vectors, and performing logical tests.

Additional topics include chaining functions, handling strings and arguments, using essential functions like *mutate*, *summarize*, and *group_by*, understanding data types, filtering rows, managing NAs, and chaining functions and datasets efficiently.

Each topic is delivered through interactive *{learnr}* package-based training to deepen users' understanding of R. These foundational concepts were selected for their relevance in dataset creation for study work. To ease the transition from SAS to R, the training aligns comparable principles between the two languages, helping programmers adapt seamlessly.

Phase 2: SDTM and ADaM

Once users are comfortable using R in a low-pressure environment, the training shifts focus to practical applications for creating Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) datasets. While SDTM and ADaM follow different standards, they share common programming fundamentals, making it effective to train them together.

This section becomes more in-depth, guiding users through real-world use cases of R in study work. Topics include working with dates and times, merging data frames with joins, pivoting data, importing external files, and performing basic data manipulation using functions like *case_when*, *case_match*, and *select().*

Training also covers R packages specific to ADaM development and how to generate datasets with metadata that align with industry standards. More advanced techniques are introduced as well, including the creation and use of custom functions.

Phase 3: Tables, Figures and Listings (TFL)

The final section of the training focuses on the implementation and creation of Tables, Figures, and Listings (TFLs) in R.

Key topics include handling special characters, working with factors, and generating figures. The module also provides comprehensive training on R packages designed to streamline TFL creation.

Additionally, users learn how to output TFLs as RTF files using techniques comparable to SAS's PROC REPORT, easing the transition for those familiar with SAS-based reporting.

Training Rollout

With a comprehensive curriculum, the R training program is designed as a long-term learning commitment spanning sixteen weeks. The schedule is divided into three main phases:

- R Basics (4 weeks)
- SDTM & ADaM (6 weeks)
- TFLs (6 weeks)

Training is structured to be self-paced, allowing users to download and complete interactive modules and exercises on a weekly basis while balancing billable work. Each week, participants meet with a dedicated R mentor and a small peer cohort to discuss questions from the self-guided content. Mentors also walk through exercises live, offering feedback and demonstrating alternative programming approaches.

To ensure true comprehension, checkpoint assessments are conducted at the end of each training phase. These may include tasks such as creating a dataset to demonstrate foundational R skills, generating and outputting an ADaM dataset with appropriate metadata, or producing a summary statistics table.

During these checkpoints, users are also encouraged to provide feedback on the interactive modules, exercises, and group sessions. This input helps ensure the training content remains relevant, clear, and easy to follow.

Training topics and exercises are predetermined for each week, based on the length and complexity of each module. Participants are expected to dedicate approximately eight hours per week to the program.

As a global organization, Syneos Health faces the unique challenge of delivering training across time zones. To support this, cohorts are divided by geographic region - North America, Europe, Asia-Pacific, and South Africa - and matched with a local mentor. Cohorts typically consist of eight to ten users, allowing for personalized interaction and support.

New cohorts are launched on a rolling basis depending on user availability. The training program is

currently open to all statistical programmers and biostatisticians within both the Full-Service Organization (FSO) and Functional Service Partnership (FSP) models. This ensures that all sponsors, regardless of service model, have access to skilled R programmers for their studies.

Challenges

Developing a scalable internal training platform came with its fair share of challenges, some expected and others surprising. Unlike typical programming tasks tied to company initiatives within biostatistics, this effort introduced unique complexities not often encountered by traditional development teams.

One of the first major hurdles was licensing for cloud services. Initially, the training modules were hosted on a standalone publishing platform, which worked well for a small group. However, as the program expanded to a broader audience, managing limited licenses became increasingly difficult. In some cases, cohorts had to be split between using desktop RStudio and the web-based platform, leading to inconsistent user experiences due to the differing interfaces.

As enrollment grew, licensing costs became a larger issue. With hundreds of potential users across a global organization, scaling on a subscription-based platform quickly became cost-prohibitive.

To address these concerns, the training team collaborated closely with the internal technology delivery engineering team. After evaluating cost, scalability, maintenance, and ease of implementation, the decision was made to transition training modules to an open-source platform with no subscription fees. This solution also offered strong, flexible security and access controls, which could be customized to meet organizational requirements.

Thanks to the engineering team's extensive experience, managing the infrastructure of an open-source solution was not a barrier, allowing the training program to scale without compromising accessibility, security, or cost-efficiency.

Future Growth and Expansion

As the R training program continues to scale, numerous opportunities for expansion and enhancement lie ahead. One key area of growth involves extending the program beyond the biostatistics teams in the Full-Service Organization (FSO) and Functional Service Partnership (FSP) models. Cross-departmental collaboration with teams such as data science, data management, and real-world/late-phase research is already being explored. This expansion would position the training platform as a unified, organization-wide resource to foster analytical excellence and support more robust, data-driven solutions for sponsors.

Additionally, the training portfolio is set to evolve with the introduction of advanced modules, tailored to further develop user capabilities. One of the most anticipated additions is training on RShiny, a powerful framework for building interactive web applications. Equipping teams with RShiny skills opens new opportunities to create dynamic dashboards and visualizations - tools that empower sponsors to make timely, data-informed decisions related to specific study endpoints and outcomes.

Another ongoing initiative involves broadening the scope of R package training. This includes both statistically intensive packages commonly used in clinical trial analysis and utility-focused packages that enhance daily programming efficiency, such as tools for streamlined data manipulation, visualization, or reproducible workflows.

By continuously expanding the curriculum and adapting to the evolving needs of the organization, the R training program aims to remain a future-focused, scalable solution - one that supports not only technical skill development but also cross-functional collaboration and innovation in clinical research.

Conclusion

As the use of R continues to grow across the clinical research industry, organizations must adapt by equipping their teams with the tools and training necessary to stay competitive, efficient, and analytically agile. The development and implementation of a scalable, interactive R training program at Syneos Health has demonstrated the value of investing in internal learning infrastructure that meets users at all levels of proficiency.

By leveraging tools like the *{learnr}* package, strategically organizing the curriculum into foundational and advanced modules, and creating a centralized access point through SharePoint, the program has successfully fostered engagement, improved skill development, and encouraged collaboration across global teams. Challenges such as licensing limitations and platform scalability were met with thoughtful, cost-effective solutions that ultimately enhanced the program's accessibility and long-term sustainability.

Looking ahead, the R training initiative is positioned for continued growth - expanding into other functional areas, incorporating advanced topics like RShiny, and deepening users' familiarity with industry-relevant R packages. Through this ongoing effort, Syneos Health aims to create a culture of continuous learning and innovation, ensuring that both current and future teams are empowered to deliver high-quality, data-driven insights for sponsors across the globe.

As the success of the R training program continues to gain visibility internally, there is growing interest from external organizations seeking to upskill their own teams in modern statistical programming practices. Given the program's structured curriculum, interactive delivery, and proven scalability, it presents a valuable opportunity for contracting as a training solution to partners and sponsors. By offering the program to those aiming to transition from SAS to R or expand their analytics capabilities - Syneos Health can support broader industry advancement while reinforcing its position as a leader in clinical data innovation. This cross-organizational collaboration has the potential to foster stronger partnerships, promote consistency in programming practices, and accelerate adoption of open-source tools across the clinical research landscape.

Acknowledgments

The authors would like to extend their sincere gratitude to the entire Biostatistics R Initiative Team for their invaluable contributions in rolling out the training program across the organization. Special thanks to Andrew Prince, Leah Crawford, Stephen Terry (Senior Statistical Programmers), and Hilary Melroy (Statistical Programmer II) of Syneos Health, whose dedication and expertise were instrumental to the success of this corporate initiative.

Contact Information

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

Michelle Page-Lopez
Syneos Health
michelle.page@syneoshealth.com
DL_Biostats_R_Team@syneoshealth.com
syneoshealth.com

Martyn Walker
Syneos Health
martyn.walker@syneoshealth.com
DL_Biostats_R_Team@syneoshealth.com
syneoshealth.com