

Piloting into the Future: Publicly available R-based Submissions to the FDA

Ben Straub, GSK

ABSTRACT

In recent years, statisticians and analysts from both industry and regulatory agencies have increased adoption of open-source software such as R. R brings great benefits from its vibrant open-source community by providing a wealth of cutting-edge statistical tools, extension packages for interactive dashboard and documentation as well as adaptability to the latest data science trends. Particularly, the dashboard focused R package Shiny has shown to provide great flexibility and interactivity. However, publicly available drug submissions with the R language as the core analysis language has been lacking and limits wider adoption within the Pharma industry. The R Consortium R Submission Working Group (RCRSWG) seeks to test the concept that a R-based language submission language can be bundled into a submission package and transferred successfully to FDA reviewers. As of May 2024, the RCRSWG has successfully completed three pilot submissions and received FDA CDER response letters. To our knowledge, these are the first publicly available submission packages that include components of open-source languages. In this talk, I will introduce the R consortium R Submission Working Group and the completed Pilot 1, 2 and 3 findings, issues that we encountered, learnings as well as current work being done in Pilot 4.

INTRODUCTION

To set the stage, I believe it is important to review the “Statistical Software Clarifying Statement” issued in May 2015 from the FDA around the use of statistical software.

FDA does not require use of any specific software for statistical analyses, and statistical software is not explicitly discussed in Title 21 of the Code of Federal Regulations [e.g., in 21CFR part 11]. However, the software package(s) used for statistical analyses should be fully documented in the submission, including version and build identification. As noted in the FDA guidance, E9 Statistical Principles for Clinical Trials (available at <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm>), “The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available.” Sponsors are encouraged to consult with FDA review teams and especially with FDA statisticians regarding the choice and suitability of statistical software packages at an early stage in the product development process.

The FDA allows pharmaceutical companies to use other software, not just specific proprietary software, for analyzing clinical drug trial data. Therefore, open-source software, like R, is permitted for a company to conduct their analysis. However, if one surveys the landscape, we see a wide chasm between the use of open-source and proprietary software for analyzing clinical trials. I believe this reluctance to accept R into a company’s toolbox is driven by several factors:

1. Option overload with R versions and packages.
2. A well-defined process for validation of open-source software.
3. Lack of publicly accessible submissions using open-source languages.

OPTION OVERLOAD

A strength of the R is its vast number of packages available to users to address specific problems. Unfortunately, this strength can also be huge hurdle to overcome as a study team or company is presented with just *too many* options to conduct an analysis. Packages sometimes don’t work well together or present challenges that newcomers to R find off-putting. Initiatives such as the tidyverse seek to develop a common API and ensure interactivity between R packages, but this is heavily scoped towards generic Data Science projects and does not give all the tools needed to conduct analysis for a drug trial. Recent initiatives like the pharmaverse, r-sassy, and openstatsware have sought to provide a

more scoped list of R packages and processes that a study team could use to produce a drug study submission. Pharmaverse and other related-initiatives have also identified gaps in the R package ecosystem for conducting drug trial analysis and have put forward additional packages and processes to help drug study teams. For example, {admiral} for ADaM datasets, {metatools/metacore} for working with dataset metadata, {oak} for SDTM datasets and many more. These external initiatives are great for helping a study team pick packages for their studies, but there is still the question of validation of R and its packages.

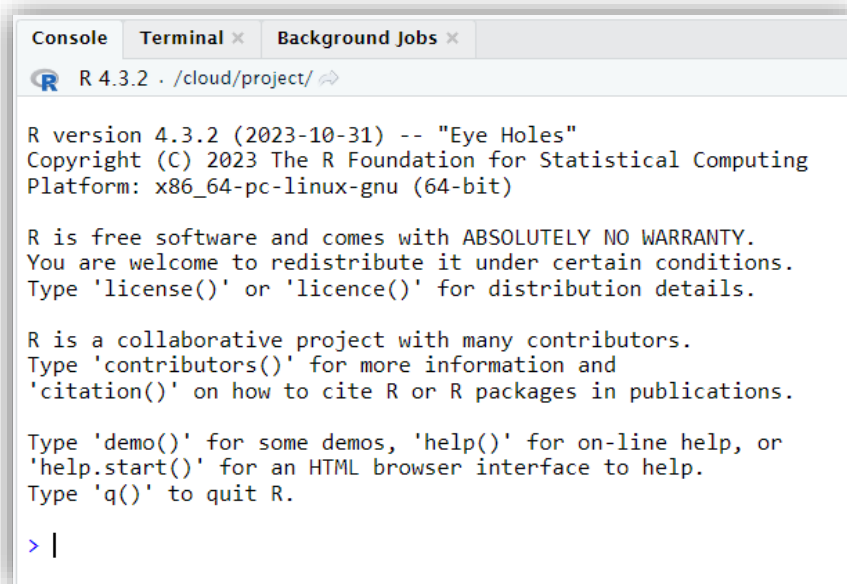
VALIDATION

A scoped list of packages dedicated to drug trial analysis is a boon to study teams and companies. However, how does a team or company ensure the integrity of the version of the packages and R being used to conduct the analysis?

Remember the FDA's Statistical Clarifying Statement:

"The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available.

R is a widely used statistical programming language with many types of analysis done over the years, but how well is it documented? When a company purchases a proprietary software, the company has strict testing and validation procedures in place to demonstrate its reliability. Documentation is readily available for the proprietary software, while for R it is harder to find on CRAN and in the source documentation. Unfortunately, when you boot up a session of R you get the following prompt in your console:



```
Console Terminal x Background Jobs x
R 4.3.2 . /cloud/project/ ↵

R version 4.3.2 (2023-10-31) -- "Eye Holes"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

This statement of “ABSOLUTELY NO WARRANTY” implies that company wishing to use open-source software must demonstrate its reliability and document its testing procedures to the FDA. The maintainers of R are not going to help you in a way that a proprietary software company might help you. With this lack of customer support, single companies or study teams might find this process overwhelming and fraught with questions on how to proceed.

Luckily, just like the above initiatives around helping study teams or companies pick R packages, there are several initiatives to help validate R and R packages available to the public. For example, pharman provides several resources on assessing package risk as well as papers and blog posts on recommendations on how to proceed around validation and common questions and concerns. The CAMIS (Comparing Analysis Method Implementations in Software) is a working group from PHUSE that

provides differences between common software used in drug trial analysis and recommendations to address. These initiatives, while still in their infancy, bring welcome resources and discussion around validation of open-source languages.

A final note on validation. A drug study can take multiple years and the versions of R and R packages can change overtime. I highly recommend establishing a process to provide study teams with a “frozen” environment. This ensures that a team’s R environment stays the same over the course of the study ensuring reproducibility and stability. In lieu of a “frozen” environment”, the R package renv can be used to provide a quasi-frozen environment. The renv package is briefly discussed later in this paper.

PUBLICLY ACCESSIBLE SUBMISSIONS

A scoped list of packages and a well-documented testing and demonstration of R’s reliability is critically important. The final challenge for further adoption of R for clinical drug trial analysis is publicly accessible R-based submissions and is the focus of this paper.

Enter the R Consortium R Submission Working Group (RCRSWG), which is a cross-industry Pharma working group focusing on improving practices of R-based clinical trial regulatory submissions. The mission of the RSWG is:

1. Easier R-based clinical trial regulatory submissions today
 - a. by showing open examples of using current submission portals
2. Easier R-based clinical trial regulatory submissions tomorrow
 - a. by collecting feedback and influencing future industry and agency decisions on system/process setup

To meet these lofty goals the RCRSWG has partnered with pharmaceutical companies and CROs to build simple publicly accessible Pilots that test the FDA’s eCTD portal for R-based submissions.

PILOT BACKGROUND

RSWG STRUCTURE

The RCRSWG is a publicly accessible group with meetings held every month between pharmaceutical, CRO, R Consortium and FDA staff. Minutes and recordings are taken and publicly available on the RCRSWG site (linked below). The focus of the public meetings is primarily report backs from the Pilot leads and feedback from the FDA on recent submissions. As of writing this paper, the meetings have been focusing on Pilot 3 with the Pilot 3 lead providing updates on the recent requested changes from FDA around the submission. The recent meetings have also had report backs on Pilot 4 as well as potential conference opportunities to present our experiences. Blockers are discussed with potential solutions provided with FDA staff and community members’ input and occasionally new potential Pilots are discussed.

As a Pilot nears completion, FDA is alerted to the use of the eCTD portal to send out the submission at these public meetings. Once received the FDA staff reviews the submission and provides the feedback both verbally, in emails as well as in official FDA letters. Questions from the FDA are generally handled during the monthly meeting, but occasionally ad-hoc meetings are set up between Pilot leads and FDA staff to discuss deeper topics.

DATA

Every pilot has made use of the publicly available Pilot CDISC data. This data was used as a Pilot to the FDA in 2006/2007 to test out the use of CDISC standards for submissions. A huge change for the industry and regulatory agencies! It seems only fitting to use the original Pilot CDISC data to investigate and document the submission of a study using R. In the appendix is a GitHub Repository containing current iteration of the Pilot data (updated in 2012) as well as a Project Report. As standards have been updated over the last twenty, the use of this older datasets has presented some interesting findings, but overall has stood the test of time!

SOFTWARE

R versions

As you probably gathered, we make use of R to do the analysis for each of the Pilots! R versions follow the semantic versioning system X.Y.Z:

- X is major version update where usually incompatible API changes are made
- Y is a minor version update with added functionality in a backward compatible manner
- Z is a patch version update to address bugs or other small issues

You can download the latest R version at CRAN (The Comprehensive R Archive Network). As of February 2024, the latest R version is 4.3.3. The pilots have been run over multiple years now and while major changes to R have not been undertaken the small changes made in the minor updates (Y) have introduced some issues for the Pilots. The biggest issue is the availability of the different R versions to FDA reviewers and employees at the various companies. Some companies have strict IT policies and requesting different versions of R can be time-consuming while other companies allow their employees to install whatever they like. The FDA also has limited availability of versions. Another issue is the Operating System that developers and reviewers use for their R versions. The FDA versions of R uses Windows while a lot of companies have Linux-based Operating Systems. This can introduce issues of file naming conventions, layout of folder/home directories, etc.

Minor versions, Y, are usually released every six months by CRAN. These R version changes have not introduced any differences in the analysis of the content in the submissions. However, different warnings around deprecation, missing dependencies as well as backwards compatibility have been noted and discussed over this multiple year project.

Base R and R packages

The version of R installed comes with many basic functions and capabilities pre-installed. However, the Base R functions are limited and can sometimes be difficult to pick up for new users as the different functions' APIs can run the gamut. The base statistical functions are also very limited, and use of specific statistical R packages are needed to conduct the analysis. The tidyverse packages are also used extensively as these provide a common API and are scoped towards data analysis as well as increasing readability of code.

In the Pilots, multiple primary packages are used to do the analysis with many more dependent packages needed for the primary package. For example, a primary package would be the tidyverse dplyr package that is used to do a lot of subsetting and manipulation of data through out all the Pilots. The dplyr package has 13 dependent packages that need to be installed alongside it. The dependent packages also have additional dependencies which can become challenging to maintain and keep a stable environment across multiple developers.

Luckily, there is an R package solution to help everyone keep a consistent and reproducible environment and minimize dependency headaches!

Renv

The stated goal of the renv package is as follows:

The renv package helps you create reproducible environments for your R projects. Use {renv} to make your R projects more isolated, portable, and reproducible.

This package was incredibly useful in helping us maintain consistent environments across developers. However, for new developers to this package and/or to R, it can be a difficult to set up and diagnose issues.

In a nutshell, {renv} works by allowing the Pilots to take a snapshot of the development environment and that snapshot is stored in a *renv.lock* file. This "lock" file stores all the information on the version of R,



version of R packages and all dependent packages and versions needed. When a developer downloads the Pilot GitHub repository and restarts their R session, the {renv} system, if not already installed will query the user to install {renv}. If {renv} is already installed, it will alert the user to packages being out of sync with the lock file. The user then has the option to install/update to the version in the lock file. If the user has everything in sync, then a everything is synced message will be displayed. Below is an example of where a user is being warned by {renv} that the lockfile is not finding packages that should be installed to conduct the analysis.

```
Restarting R session...

* Project '/cloud/project' loaded. [renv 0.17.0]
Current project R package repository:
  https://packagemanager.posit.co/cran/2023-03-15

Below R package path are searching in order to find installed R packages in this R session:
  /cloud/home/r590548/.cache/R/renv/library/project-4b716941/R-4.2/x86_64-pc-linux-gnu
  /cloud/project/renv/sandbox/R-4.2/x86_64-pc-linux-gnu/e11edd0e

The project home directory is /cloud/project

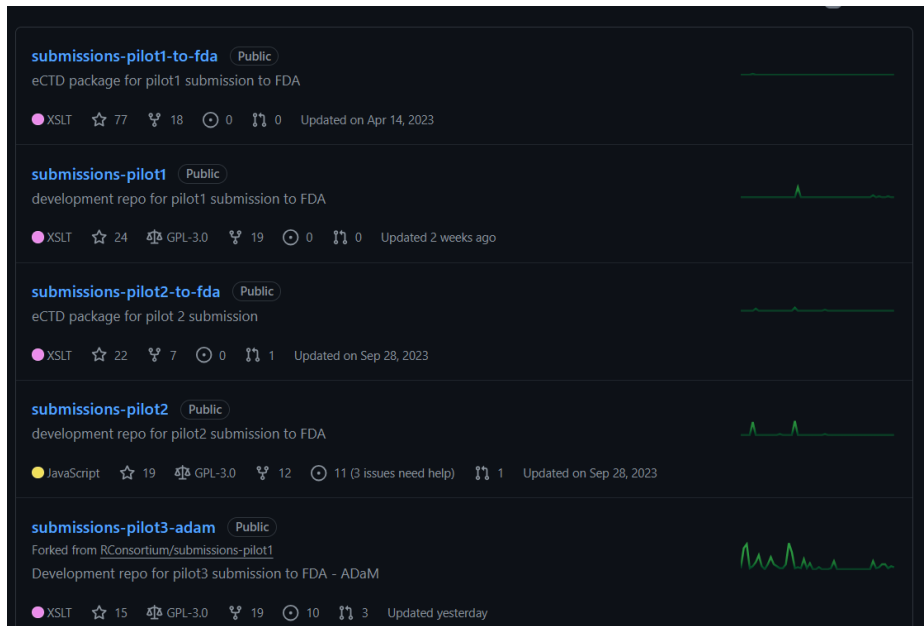
* One or more packages recorded in the lockfile are not installed.
* Use `renv::status()` for more details.
> |
```

Github/git

We make use of GitHub/git to document changes within the codebase being used in each of the Pilot submissions. GitHub is the web-based interface to the version-control system git. GitHub allows the Pilots to publicly discuss in Pull Requests, a procedure for requesting changes to the code, in the open public. Issues can be created to address topics or concerns that are also publicly available. The Issues can be tied to Pull Requests and closed automatically if the updated code meets the requirements in the Issue. Occasionally, the Issues are long-form discussion that span many months. Most Pilots close their Issues upon successful completion, but all are still publicly available.



The typical GitHub structure of the Pilots consists of two repos specific to each Pilot. The first repo is the development repo where the code changes and outputs are produced. The second repo is preparation for submission to the eCTD portal and have the suffix "-fda". This repo has a pseudo representation of some of the modules needed to do submissions, e.g. m3 or the m5 modules. In the screenshot below you can see Pilots 1-3 development and submission repositories.



eCTD portal

The final piece of software that the Pilots were intimately familiar with was the eCTD portal and its supporting documents. The “Electronic Common Technical Document” is the standards put in place by the FDA for submissions. Documents around submission specifications should be noted, for example “Specifications for File Format Types Using eCTD Specifications”, which is linked below. Please note that this document has the `.r`` file as allowed to be submitted in the M3-M5 modules! A standard R convention is to end the R script with a `.R` rather than a `.r`. This has no impact on the actual file, but should be noted that the eCTD file does not accept `.R` files. Another important document to become familiar with is the “Providing Regulatory Submissions in Electronic Format — Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications”, which is linked below as well.

PILOT 1 (4 DISPLAYS)

Initial work for this Pilot began in 2020 with a focus on delivering 4 displays to FDA via the eCTD portal. I would like to take this time to note the paucity of displays in the submission – just 4!. The purpose of the Pilots is to test the mechanisms to deliver an R-based submission to the FDA and not reproduce a large body of work in R, hence the small number of displays.

The displays consist of 3 tables and 1 figure. The 3 tables should be noted in how they were rendered – rtf and pdf. While the number of displays created were limited, the number of functions, helper files and a helper package needed to support the analysis quickly ballooned! A helper package called `{pklite}` was built alongside the Pilot 1 submission to assist in “packaging” up the materials for submission via the eCTD portal. You can read more about the `pklite` process at [R for Clinical Study Reports and Submission \(r4csr.org\)](#). To give a “simple” one line statement of `pklite` – it takes the entire R submission and makes a `.txt` file that can be submitted through the eCTD Portal and using `{pklite}` the FDA reviewer can quickly rebuild everything needed.

Pilot 1 was officially submitted on November 22, 2021. This submission was the “first publicly available” R-based submission to the FDA. The submission included an R package, R scripts for analysis, R-based analysis data reviewed guide (ADRG), and other important components. The final response letter from the FDA was received on March 14, 2022.

This Pilot has successfully concluded.

PILOT 2 (4 DISPLAYS IN SHINY)

This is one of the “first publicly” available R-based submission packages that contains a Shiny application. Shiny is an R package that provides “elegant and powerful web framework for building web applications.” How could this help with drug trial analysis submissions? The ability to “drill-down” into data is at the heart of any drug trial analysis. Trying to understand the interactions of a drug with a patients’ safety profile to their chemistry interactions on down to the different sub-groups analysis is critical to a successful submission. Unfortunately, most of the content doing this analysis is static. If additional requests are needed, then additional static content must be produced. Shiny-based submissions could allow for greater drill-down capability and reduce additional requests for content and hopefully speed up submission timelines

The main goal of Pilot 3 was to test the submission of an R-based Shiny application bundled into a submission package and transfer it successfully to FDA reviewers. The submitted application was built using the datasets and analyses that were used for the R Submission Pilot 1. The four displays created in the Pilot 1 were now bundled into a single Shiny App. The Pilot 2 team made use of the pkglite package to help bundle the necessary files to deliver to the FDA reviewers via the eCTD portal. The FDA reviewers were able to successfully deploy the Shiny App and review the data. A deployed version of the Shiny Pilot 2 application is available on the Appsilon Blog (linked below). The final response letter from the FDA was reviewed on September 27, 2023.

In this submission, there were many open-source R packages that were used to create and execute the Shiny application. A very well-known shiny-based interactive exploration framework {teal} was used mainly for analyzing the clinical trial data. The full list of open-source and proprietary R analysis packages is available on this Analysis Data Reviewer’s Guide prepared by the R Consortium R Submissions Working Group for the Pilot 2.

This Pilot has successfully concluded.

PILOT 3 (4 DISPLAYS AND 5 ADAMS)

Pilot 3 differs from Pilot 1 and Pilot 2 in that the ADaMs used in the displays are now built using R. The previous two pilots had used the original CDISC Pilot data ADaMs and did not reproduce them in R. As ADaM datasets have associated metadata/specification files with them additional packages were used to apply the metadata called metatools/metacore and xportr. Therefore, the Pilot not only explored building ADaMs and Displays with R code, but also how to work with the associated metadata data. The Pilot 3 team only produced 5 ADaM datasets, which were the 5 ADaMs needed for the displays. The 5 ADaMs were:

1. ADSL – Subject-Level Analysis Dataset
2. ADADAS – ADAS-COG Analysis Dataset
3. ADAE – Adverse Events Analysis Dataset
4. ADLBC – Analysis Dataset Lab Blood Chemistry
5. ADTTE – AE Time to 1st Derm. Event Analysis

It should be noted again the limited number of ADaMs produced, which only pertained to the ADaMs needed to re-produce the displays and to only test the mechanisms to deliver an R-based submission to the FDA via the eCTD portal. Pilot 3 documented Findings on their GitHub wiki where differences in the original data and the newly created ADaMs in R were found. None of these findings impacted the displays, but I do encourage readers to review the findings on Pilot 3 Github repo (linked below).

This pilot was successfully submitted to the FDA on Aug 28, 2023. This is the “first publicly available” R submission that included R scripts to produce ADaM datasets and TLFs. Pilot 3 also differs from Pilot 1 and Pilot 2 in that {pkglite} was not used to deliver the submission package. Instead, all code was placed into the M3-M5 modules successfully and submitted to the FDA. A helper package was used in the

submission and the FDA was able to successfully download the package from GitHub. However, this was frowned upon as a potential practice for future submission. It was decided to do a re-submission of Pilot 3 exploring a way to deliver the same package as a .zip file through the eCTD portal.

This Pilot has a re-submission planned for early April 2024.

PILOT 4 (WEBASSEMBLY AND CONTAINERS)

This pilot aims to explore using technologies such as containers and WebAssembly software to package a Shiny application into a self-contained unit, streamlining the transfer and execution process for enhanced efficiency.

This pilot is expected to be divided into two parallel submissions the first part will investigate WebAssembly and the second part will investigate containers. It should be noted that two of the developers for Pilot 4 are presenting at Pharmasug 2024 with the paper: “Experimenting with Containers and webR for Submissions to FDA in the Pilot 4” and I highly recommend you check that out!

This Pilot is still in its planning and exploration stages.

CONCLUSION

R is a powerful open-source language with a lot of desirable features for conducting analysis of drug trial data. The wider adoption of the language for submission is stymied by the complicated nature of R and package versions, the wide array of packages, the need to show validation, and the paucity of publicly available submissions. Multiple initiatives over the years have sprouted to help address these challenges. In this paper, I have shown how the R Consortium R Submission R Working Group have developed simple and “publicly available” Pilots to test and document submitting an R-based submission to the FDA.

REFERENCES

R Consortium R Submission Work Group. Accessed March 23, 2024

<https://rconsortium.github.io/submissions-wg/>

[RConsortium/submissions-pilot3-adam: Development repo for pilot3 submission to FDA - ADaM \(github.com\)](https://github.com/RConsortium/submissions-pilot3-adam). Accessed on March 23rd, 2024. <https://github.com/RConsortium/submissions-pilot3-adam>

[RConsortium/submissions-pilot2: development repo for pilot2 submission to FDA \(github.com\)](https://github.com/RConsortium/submissions-pilot2). Accessed on March 23rd, 2024. <https://github.com/RConsortium/submissions-pilot2>

[RConsortium/submissions-pilot1: development repo for pilot1 submission to FDA \(github.com\)](https://github.com/RConsortium/submissions-pilot1). Accessed on March 23rd, 2024. <https://github.com/RConsortium/submissions-pilot1>.

[RConsortium/submissions-pilot4: Submissions WG Pilot 4 \(github.com\)](https://github.com/RConsortium/submissions-pilot4). Accessed on March 23rd, 2024. <https://github.com/RConsortium/submissions-pilot4>.

FDA US Food & Drug Administration. “Statistical Software Clarifying Statement.” Accessed February 15, 2024. <https://www.fda.gov/media/161196/download> [Statistical Software Clarifying Statement \(fda.gov\)](https://www.fda.gov/media/161196/download).

Pilot 2 Shiny App. [Advancing FDA Clinical Trial Submissions with R: Reproducing the R Submissions Pilot 2 Shiny App Using Rhino \(appsilon.com\)](https://www.appsilon.com/post/fda-clinical-trial-submissions-with-r-shiny-rhino) Accessed March 23, 2024 <https://www.appsilon.com/post/fda-clinical-trial-submissions-with-r-shiny-rhino>.

CDISC Pilot Project. Accessed February 28, 2024. <https://github.com/cdisc-org/sdtm-adam-pilot-project>

[openstatsware](https://www.openstatsware.org/) - scientific working group of the American Statistical Association (ASA) Biopharmaceutical section (BIOP). Accessed on March 23, 2024. <https://www.openstatsware.org/>

[pharmaverse](https://pharmaverse.org/) - A connected network of companies and individuals working to promote collaborative development of curated open source R packages for clinical reporting usage in pharma. Accessed on March 23, 2024. <https://pharmaverse.org/>

pharmar.org - R Validation Hub is a collaboration to support the adoption of R within a biopharmaceutical regulatory setting. Accessed on March 23, 2024.

[CAMIS - A PHUSE DVOST Working Group \(psiaims.github.io\)](https://psiaims.github.io). Accessed on March 23, 2024.
<https://psiaims.github.io/CAMIS/>

[CRAN - The Comprehensive R Archive Network \(r-project.org\)](https://cran.r-project.org/). Accessed on March 23, 2024.
<https://cran.r-project.org/>

[Specification for Transmitting Electronic Submissions using eCTD Backbone Specifications \(fda.gov\)](https://www.fda.gov/media/85816/download). Accessed on March 23, 2024. <https://www.fda.gov/media/85816/download>

[Providing Regulatory Submissions in Electronic Format — Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications Guidance for Industry \(fda.gov\)](https://www.fda.gov/media/135373/download?attachment). Accessed on March 23, 2024. <https://www.fda.gov/media/135373/download?attachment>

[Pilot 3 QC Findings · RConsortium/submissions-pilot3-adam Wiki \(github.com\)](https://github.com/RConsortium/submissions-pilot3-adam/wiki/QC-Findings). Accessed March 23, 2024. <https://github.com/RConsortium/submissions-pilot3-adam/wiki/QC-Findings>

ACKNOWLEDGMENTS

GSK and my manager, Vinitha Arumugam, for allowing me to explore and use open-source tools.

RECOMMENDED READING

- [R/Adoption Series: R and Shiny in Regulatory Submissions Webinar - R Consortium \(r-consortium.org\)](https://www.r-consortium.org/r-adoption-series-r-and-shiny-in-regulatory-submissions-webinar) - <https://www.r-consortium.org/r-adoption-series-r-and-shiny-in-regulatory-submissions-webinar>
- [PRE_London09.pdf \(phuse.s3.eu-central-1.amazonaws.com\)Project Environments • reny \(rstudio.github.io\)](https://phuse.s3.eu-central-1.amazonaws.com/Project%20Environments%20renv/rstudio.github.io)
- [Novo Nordisk's Journey to an R based FDA Submission - YouTube](https://www.youtube.com/watch?v=t33dS17QHUA). Accessed March 23, 2024. <https://www.youtube.com/watch?v=t33dS17QHUA>

CONTACT INFORMATION

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

Ben Straub
GSK
ben.x.straub@gsk.com

Any brand and product names are trademarks of their respective companies.