PharmaSUG 2021 - Paper SI-108

Blogdown and Bookdown: Using R packages to document and communicate new processes to Clinical Programming

Ben Straub, GlaxoSmithKline

ABSTRACT

The landscape of Clinical Reporting is changing fast. Open source languages that used to be off-limits or seen as only niche hobbies are now being embraced industry wide. Providing documentation and communications surrounding the use of an open source language can be a herculean task for any Clinical Programming department. At GSK, the Clinical Programming (GSK-CP) department has decided to embrace the use of the R language as its open source language of choice. The choice of R has allowed the GSK-CP department to make great use of two packages, bookdown and blogdown, to help document and communicate the use of R for Clinical Reporting. This paper will focus on GSK-CP's journey with using R and how bookdown and blogdown have helped to accelerate and codify our use of R for Clinical Reporting. Readers will be given a brief background on how R was initially adopted within GSK-CP, issues surrounding R that were identified and solved, processes that were established to help promote Good X Practices (GXP) with R and how we utilized bookdown and blogdown to make it all happen.

INTRODUCTION

Historically, clinical research and pharmaceutical drug development have relied heavily on the SAS programming language for database transformations and generation of analysis displays for regulatory submissions. Recently, the industry has witnessed a growing interest in open source languages such as R and Python as an alternative to SAS for many activities related to clinical research. Currently, the Clinical Programming department within GSK-Biostatistics is leading an effort to implement the use of R in their Clinical Reporting pipeline. The effort to increase the use of R within GSK-Clinical Programming (GSK-CP) is not to replace SAS, but to supplement options for programmers, accelerate onboarding of new hires and stay current with industry standards. For example, a concerning trend is that many new hires from universities are more proficient in R than in SAS. Spending resources and time on teaching SAS to new hires is expensive and does not help to accelerate studies pipelines. A new hire being able to jump into display creation using R would be a boon for any study team.

GSK-CP is not alone in the industry with this effort to make use of R. Groups such as the R Consortium, R/Pharma, and RStudio are heavily invested in figuring out common problems and publicizing best practices. These active groups also see involvement from regulatory agencies such as the FDA. However, every company will have a bespoke implementation of R due to resources, internal SOPs, skill levels and other factors. These bespoke implementations will give rise to the need for internal documentation. Sometimes the ideal solution for one company might not work in another company. Within GSK-CP, we have made great use of bookdown and blogdown to help document and educate on our recommended best practices. Bookdown and blogdown are R packages that allow R programmers to collate and present documentation and code in visually appealing, elegant and interactive ways.

Finally, I believe that sharing our story helps to engender confidence at other companies and demystify the use of R within a regulated environment. As stated before, the road being travelled is in good company!

NOTE: Technical discussion is limited here. A companion paper*, Blogdown and Bookdown: Deep dive of the R packages and components needed to create documentation and* website is available in the Advanced Programming section for more discussion on RMarkdown, bookdown and blogdown.

R ADOPTION

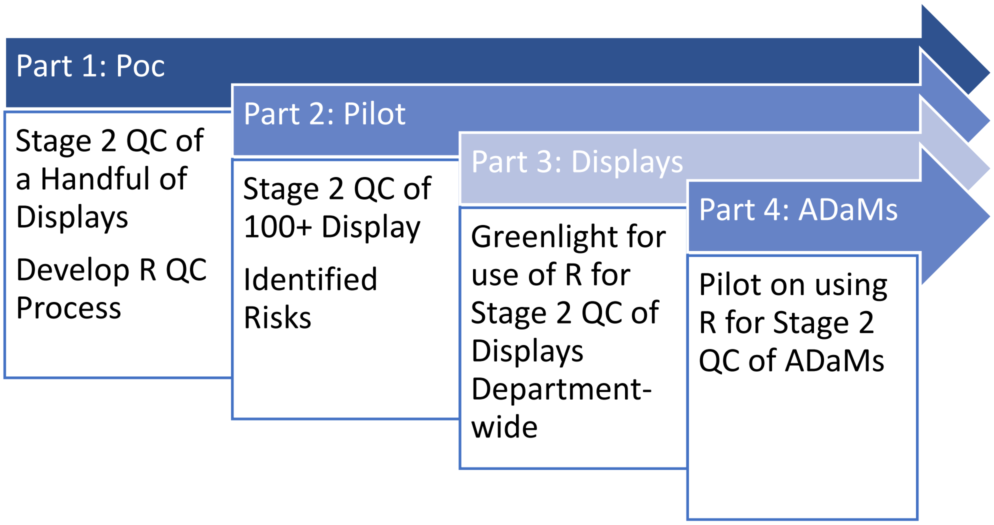
Before I begin proselytizing on the many amazing possibilities of blogdown and bookdown I would like to set the stage for how R was adopted into the Clinical Reporting process at GSK-CP.

The back story will help give context on the following:

- The use of R within GSK-CP in the summer of 2019
- How R became adopted within Clinical Programming at GSK
- The need for extensive documentation versus Quick Starts versus flare/propaganda/blog posts
- Discovery of bookdown and blogdown from other internal groups at GSK
- How we made use of the many features to help codify and disseminate R

THE BACK STORY

The adoption of R so far at GSK-CP can be broken down into four large parts – Proof of Concept, a Shadow Pilot, R4QC of Displays and R4QC of ADaM data sets. Each of the four parts helps to inform the decisions and process of the next part. Each part also identified issues that were native to R and needed to be addressed. These issues are interesting to Clinical Reporting and I make a point to highlight those issues within each section. The tools that GSK-CP utilized in the beginning also play a big role in how we adopted the use of bookdown and blogdown. These tools are Git/GitHub, a type of version control software and RStudio an Interactive Development Environment (IDE) that is primarily focused on R, and the suite of R Packages called the tidyverse. Please note that GSK-CP is still in the process of adopting R more widely within the department.


Part 1: Poc

Stage 2 QC of a Handful of Displays

Develop R QC Process

Part 2: Pilot

Stage 2 QC of 100+ Display

Identified Risks

Part 3: Displays

Greenlight for use of R for Stage 2 QC of Displays Department-wide

Part 4: ADaMs

Pilot on using R for Stage 2 QC of ADaMs

Figure 1: Parts within Parts
Note: Stage 2 QC is sometimes referred to as Double or Independent Programming.

NOTE: GSK-CP is closely aligned with an enterprising and R-centric group called Statistical Data Sciences (SDS). The first year we relied heavily on them to help us with using R. SDS is comprised of mostly Data Scientists, but they do have clinical programmer and statisticians embedded within their team. Most of the staff within SDS have participated in some form of the Clinical Reporting process in the past.

.

The R Landscape within GSK-CP in Summer of 2019

R is heavily used at GSK within many departments, but in 2019 it was not widely used within GSK-CP. Several studies had made use of the package ggplot2 for developing some enhanced graphics and others had used it for exploratory purpose, but guidance on its use was limited and most of GSK-CP was comprised of SAS programmers. Efforts were a foot to promote the use or R from leadership with an eye on new hires and staying current with industry. To address this issue, SDS developed several trainings to help onboard staff to the world of R and the suite of packages called Tidyverse. These trainings were delivered in the summer of 2019 to staff and helped to propel us forward.

Part 1 – Proof of Concept

In the summer of 2019, within GSK-CP, a small team of Clinical Programmers met to discuss the possibility of utilizing R for Clinical Reporting. The small group of CPs had limited knowledge of using R – most had some experience within university setting, but all were SAS programmers. The success of the internal training on R from SDS led this group to believe that R could be utilized. This small team came up with a small Proof of Concept to demonstrate to the CP Leadership and fellow programmers that R was a viable language for Clinical Reporting.

This initial foray into R for Clinical Reporting was small and concentrated. We decided upon a small set of displays within an already completed Study to do double programming/Stage 2 QC on. The team felt this was a low risk situation and would help programmers become familiar with R in GSK's ecosystem. Staff within SDS had already built a RStudio Server Environment with the help of RStudio. The RStudio Server interacted seamlessly with our study data just like our SAS servers. We just had to take the plunge. Matching the production display programs with R proved very straightforward. The displays were mostly counts, percents and simple summary statistics. However, we did identify some key issues during the proof of concept.

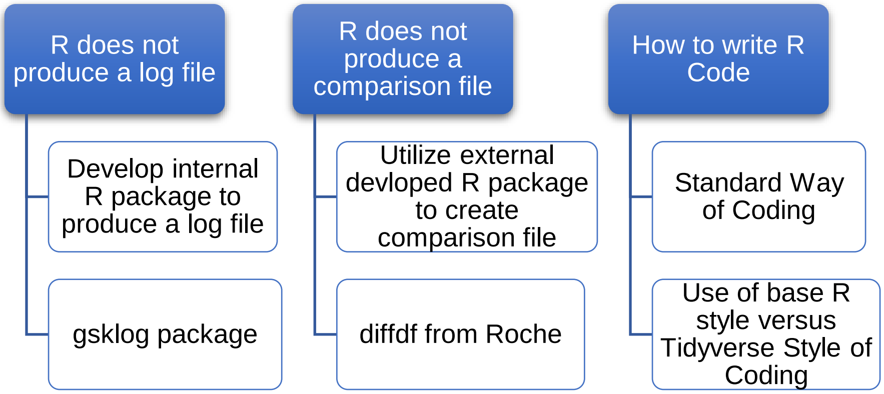

R does not produce a log file
Develop internal R package to produce a log file
gsklog package
R does not produce a comparison file
Utilize external devloped R package to create comparison file
diffdf from Roche
How to write R Code
Standard Way of Coding
Use of base R style versus Tidyverse Style of Coding

Figure 2: Issues and solutions identified from Proof of Concept

Part 2 – Shadow Pilot

Early fall of 2019, the R4QC team was given the greenlight to push forward with a Shadow Pilot Study. It was decided that the R4QC team was to mass Stage 2 QC close to 100+ displays during an ongoing study. We were not the official QC programmers, hence "shadow", as it was deemed a risk to let a team of relatively inexperienced R users do formal Stage 2 QC of display. However, the team thought this effort was important to shadow an ongoing study as it gave a real feel of doing actual study work within a tight timeline.

Several issues were quickly brought to the forefront that did not show up in the Proof of Concept. For example, the team of programmers were joined by a group of statisticians to write QC code for efficacy displays. GSK SOPs states that Statisticians most do both production and QC of any display that have modeling involved. It was discovered that some of the underlying assumptions behind R and SAS modeling algorithms differ slightly. These assumptions can cause slight differences in matching production to qc outputs. Also, along the way we found that the R and SAS differ in their rounding assumptions. Finally, we found that the package diffdf, while incredibly helpful for doing comparison of production and qc outputs, did not give us the needed abilities to document results from comparisons.

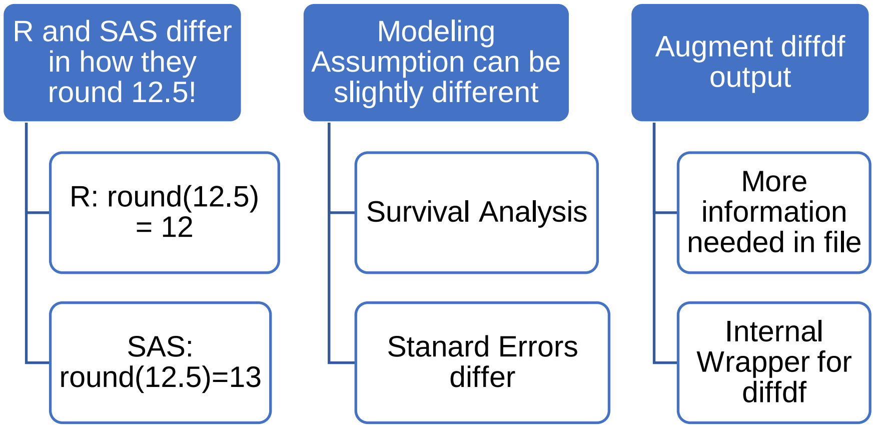

R and SAS differ in how they round 12.5!
R: round(12.5) = 12
SAS: round(12.5)=13
Modeling Assumption can be slightly different
Survival Analysis
Stanard Errors differ
Augment diffdf output
More information needed in file
Internal Wrapper for diffdf

Figure 3: Issues identified from Shadow Pilot

Part 3 – R4QC

The pilot, while identifying several issues, was deemed a success by the CP-Leadership Team and all those involved. We were very excited to unleash R to the masses! At first, the R4QC team thought the best path forward was to identify a study and utilize R for Stage 2 QC exclusively. Unfortunately, study timelines being so tight, it was difficult to identify one study within our timelines and persuade teams to jump on board. Study Lead Programmers were nervous that staff resources would be rate-limiting, i.e. what happens if all programmers on the study only know SAS or have minimal programming and minimal interest in using R. Would their study be put at risk?

Therefore, it was deemed useful to open the flood gates and allow all Clinical Programmers for any study to engage the use R in their work on Stage 2 QC of displays. Here programmers for any study could engage Stage 2 QC with R. While the gates were technically wide open, they were heavily policed. Programmers were expected to engage with the R4QC team as well as their managers and Study Lead Programmers to declare their use of R. The R4QC team also needed document who and where the programs were being developed. We turned to the ever tried and true method of tracking outputs in Excel. However, we also needed to develop internal resources that showcased GPP, code snippets, process flows within CP. Enter blogdown and bookdown as a single source of truth for using R for Stage 2 QC.

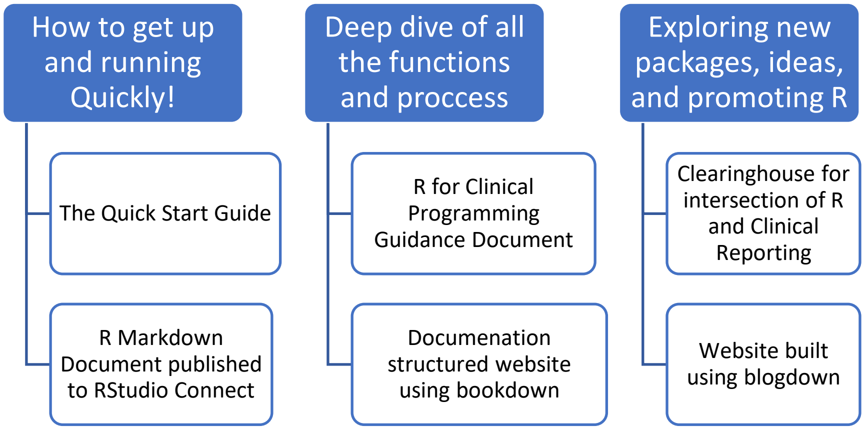

How to get up and running Quickly!
The Quick Start Guide
R Markdown Document published to RStudio Connect
Deep dive of all the functions and proccess
R for Clinical Programming Guidance Document
Documenation structured website using bookdown
Exploring new packages, ideas, and promoting R
Clearinghouse for intersection of R and Clinical Reporting
Website built using blogdown

Figure 4: Enter the downs

Part 4 – ADaM QC

In the fall of 2020, a second R pilot project was launched that explored the use of R for data set programming. This was identified as a needed addition to our suite of abilities of QC display programming. Most programmers spend a lot of their time programming ADaM data sets. The individuals involved with this part of the programming found R to match SAS' abilities for data set programming.

With data set programming in R being identified as low risk it was decided to enhance the infrastructure that was built for R4QC of displays.

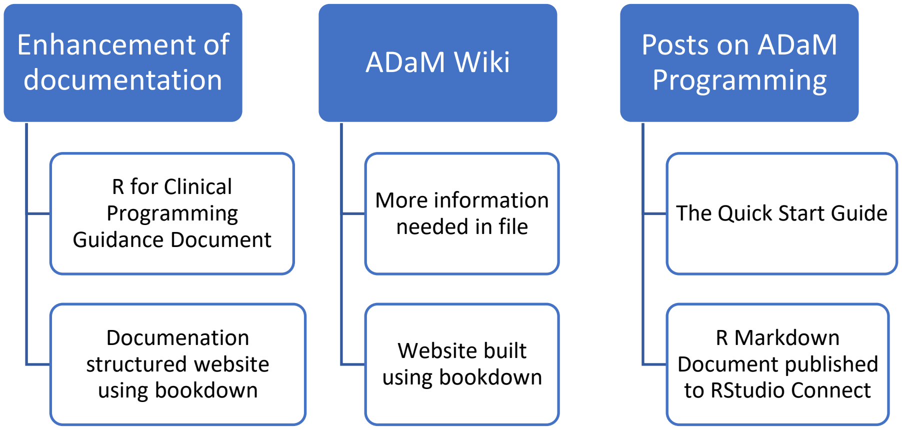

Enhancement of documentation
R for Clinical Programming Guidance Document
Documenation structured website using bookdown
ADaM Wiki
More information needed in file
Website built using bookdown
Posts on ADaM Programming
The Quick Start Guide
R Markdown Document published to RStudio Connect

Figure 5: Enhancing the downs with ADaMs

BUILDING DOCUMENTATION AND WEBSITES

The team leading this effort has made heavy use of the R packages bookdown and blogdown to both communicate guidelines and best practices as well showcase specific use-cases. Before we go into detail on how GSK-CP strategically implemented the use of bookdown and blogdown it will be wise for us to discuss some important features of the R Ecosystem:

1) **What is R and RStudio?** R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. RStudio is an Interactive Development Environment (IDE) that sits on top of R. RStudio has many nice quality-of-life improvements that makes using R much easier. Some of those features

allow blogdown and bookdown to render websites easily. In my opinion, I think R's original design being focused on statistical analysis is much more akin to SAS than say Python. I love Python, but it is so vast and was developed for a different purpose other than statistical analysis.

2) What is an R Package? The capabilities of R are extended through user-created packages, which allow for many specialized techniques, as you have gathered, bookdown and blogdown, are R packages that help develop websites. The beauty of these two packages is that they do a lot of the heavy lifting for you in developing your website.

3) What is Github? Github is a website that allows for a cloud-based Git repository hosting service. Git is a type of version control software and Github helps to improve the Git software with additional quality of life enhancements

The relationship between R, RStudio and GitHub is important. While R can be done in the command-line, RStudio can make your life immensely easier with interacting with Packages and Github/Git. We would not have been able to build our documentation and websites without these three components interacting seamlessly.

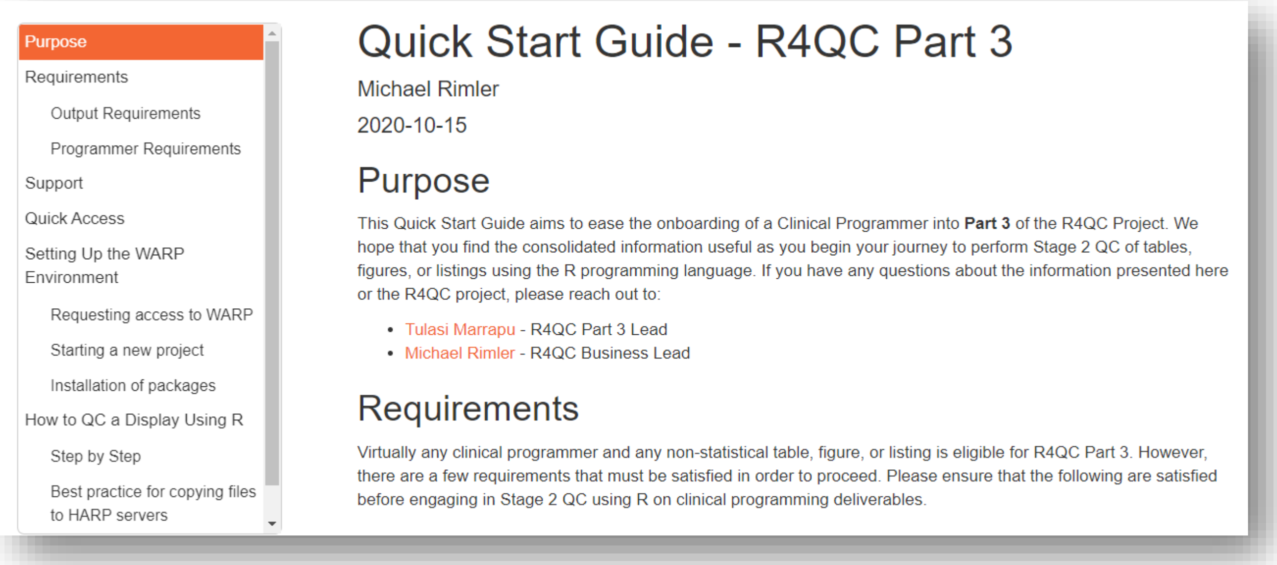
BOOKDOWN

During the discussion of Part 1- POC several programmers on that team took an internal training course called *Introduction to R and Tidyverse* that was created by the SDS team. The material developed was in the context of Clinical Reporting, but on a high-level, i.e. they did not get into the minutiae of double/independent programming, creating log and comparison files, rounding in SAS and R, alignment of columns, etc. However, the SDS team had made available the source code for their training on the internal GSK github. This is an important part! The code was freely available for others to use internally in a public repository. Everything needed to build a new bookdown was freely and readily available. We, the R4QC team, just had to adopt the training materials to our specific needs.

As stated before, bookdown is an R package. Essentially bookdown takes a series of documents that contain syntax and code and creates a book. The book can then be published to a website or shared with a group. At GSK, we make use of RStudio Connect Pro to publish our documents. You can publish to other hosting service, e.g. netifly.

A Tome versus a Quick-Start

Briefly mentioned in Part 3, there was an identified need to get programmers up and running quickly. While the importance of epic documentation is important, we also did not want to scare away folks. Therefore, a Quick Start Guide was born. The Quick Start gives background on how to get access to R, acceptable uses of R and a brief overview of the big steps needs to complete Stage 2 QC of a display. The document is developed using RMarkdown and published to RStudio Connect. Users can bookmark the site and refer to it as needed. The publishing process is important to note here. We have another server that we can publish websites to – either simple RMarkdown documents, documentation or interactive websites. Also, please note the use of RMarkdown, which forms the backbone of bookdown and blogdown – please see companion paper for discussion on RMarkdown.



Quick Start Guide - R4QC Part 3

Michael Rimler

2020-10-15

Purpose

This Quick Start Guide aims to ease the onboarding of a Clinical Programmer into **Part 3** of the R4QC Project. We hope that you find the consolidated information useful as you begin your journey to perform Stage 2 QC of tables, figures, or listings using the R programming language. If you have any questions about the information presented here or the R4QC project, please reach out to:

- Tulasi Marrapu - R4QC Part 3 Lead
- Michael Rimler - R4QC Business Lead

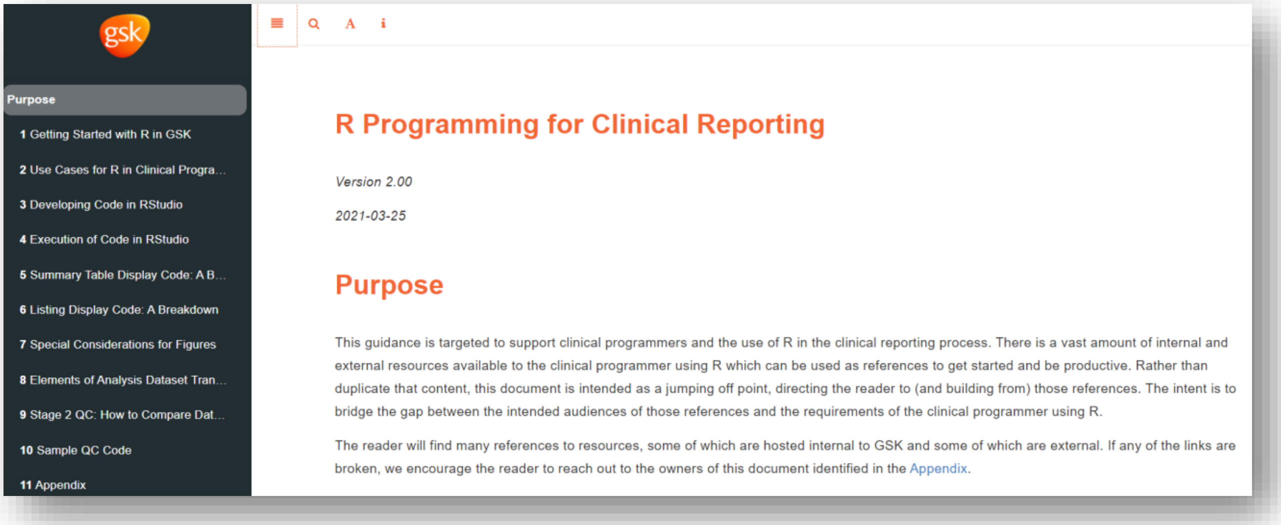
Requirements

Virtually any clinical programmer and any non-statistical table, figure, or listing is eligible for R4QC Part 3. However, there are a few requirements that must be satisfied in order to proceed. Please ensure that the following are satisfied before engaging in Stage 2 QC using R on clinical programming deliverables.

Figure 6: Get up and running fast!

Design of our R Guidance Documentation

Embedding code into a document can be challenging. Screenshots can be taken to help keep the formatting or you can mess around with code formatting within the document. However, what if you discover a bug in your code or develop new arguments/parameters for the functions. You now must take new screenshots and insert into the document. Bookdown makes this seamless. You can automatically update the function or fix that bug and republish. This also forms a bit of informal testing of your new features. Let's say you have a package that has been updated and breaks your code within the bookdown guidance document. While this is unfortunate, it also reminds you to address your code before it goes out to the masses!


gsk
Purpose
1 Getting Started with R in GSK
2 Use Cases for R in Clinical Progra...
3 Developing Code in RStudio
4 Execution of Code in RStudio
5 Summary Table Display Code: A B...
6 Listing Display Code: A Breakdown
7 Special Considerations for Figures
8 Elements of Analysis Dataset Tran...
9 Stage 2 QC: How to Compare Dat...
10 Sample QC Code
11 Appendix
A

R Programming for Clinical Reporting

Version 2.00

2021-03-25

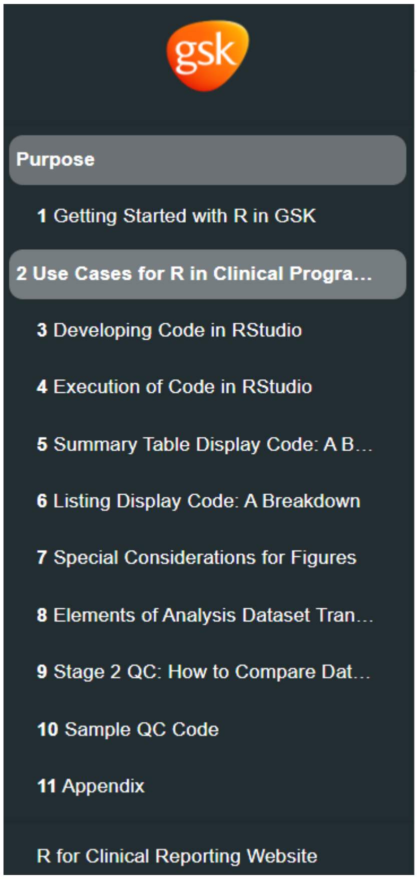
Purpose

This guidance is targeted to support clinical programmers and the use of R in the clinical reporting process. There is a vast amount of internal and external resources available to the clinical programmer using R which can be used as references to get started and be productive. Rather than duplicate that content, this document is intended as a jumping off point, directing the reader to (and building from) those references. The intent is to bridge the gap between the intended audiences of those references and the requirements of the clinical programmer using R.

The reader will find many references to resources, some of which are hosted internal to GSK and some of which are external. If any of the links are broken, we encourage the reader to reach out to the owners of this document identified in the Appendix.

Figure 7: Home page of R Programming for Clinical Reporting created using bookdown

The current iteration of our R Programming for Clinical Reporting Guidance document has 10 chapters with a short Appendix. Below is side panel of the guidance documentation showing the table of contents. I briefly discuss the sections of below, but each section is not brief. The abilities of bookdown is just like any website. We have embedded videos, sample code, interactive tables within each section.



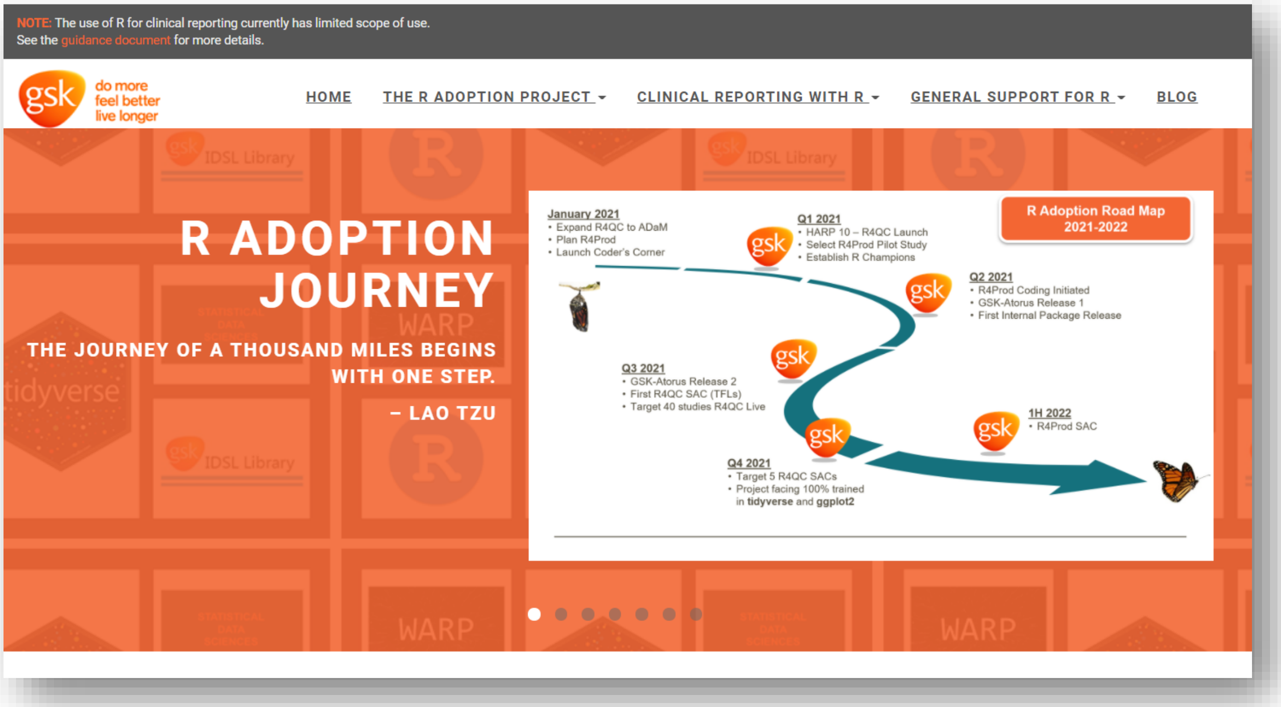
- The first 4 chapters deal with the high-level overview of using R within Clinical Programming at GSK. These Chapters focus on what we can currently use R for, no production work yet, and how to insert headers, execute logs and archive files within our GSK systems.
- Chapters 5-6 deal with actual execution of the code for display programming. Chapters 5 and 6 are deep dive into pieces of code that follow closely to displays that could be created in a real study.
- Chapter 8 focuses on ADaM Programming. We found code snippets for executing certain common tasks, ADT calculation or ANL01FL assignment to be more ideal than code walkthroughs.
- Chapter 9 showcases how to use comparison tools from the diffdf package. Chapter 9 also has documentation on SAS versus R differences and how to address them.
- In Chapter 10, we have all the sample Code ready to be downloaded and used that was discussed throughout the document.
- I have a publicly available sample site that closely mimics our guidance document. It is also discussed within the companion paper.

BLOGDOWN

Documentation is important to any process, especially in a regulatory environment. The bookdown is a standalone website that a user can access easily and read through the materials. It is possible to spruce up the bookdown site to make it a little more dynamic and exciting to visit. However, this sprucing up is unnecessary when you unleash the power of the blogdown package.

The blogdown package essentially takes website development and puts it into the hands of a R Programmer. With just surface-level knowledge of web development, you can pull off a decent and interactive website. As clinical programmers at GSK, we were the target audience for the design of blogdown, i.e. limited knowledge of web development, but now very familiar with R and RStudio Luckily, our friends within SDS had also developed a website using blogdown. The enterprising folks at SDS had again made their source code available on GSK's github in a public repositor. The R4QC team was able to initialize a website and repurpose to serve our initiative – i.e. increasing the adoption of R within

Clinical Programming.



NOTE: The use of R for clinical reporting currently has limited scope of use.
See the guidance document for more details.

R ADOPTION JOURNEY

THE JOURNEY OF A THOUSAND MILES BEGINS WITH ONE STEP.

– LAO TZU

R Adoption Road Map 2021-2022

January 2021
- Expand R4QC to ADaM
- Plan R4Prod
- Launch Coder's Corner

Q1 2021
- HARP 10 – R4QC Launch
- Select R4Prod Pilot Study
- Establish R Champions

Q2 2021
- R4Prod Coding Initiated
- GSK-Atorus Release 1
- First Internal Package Release

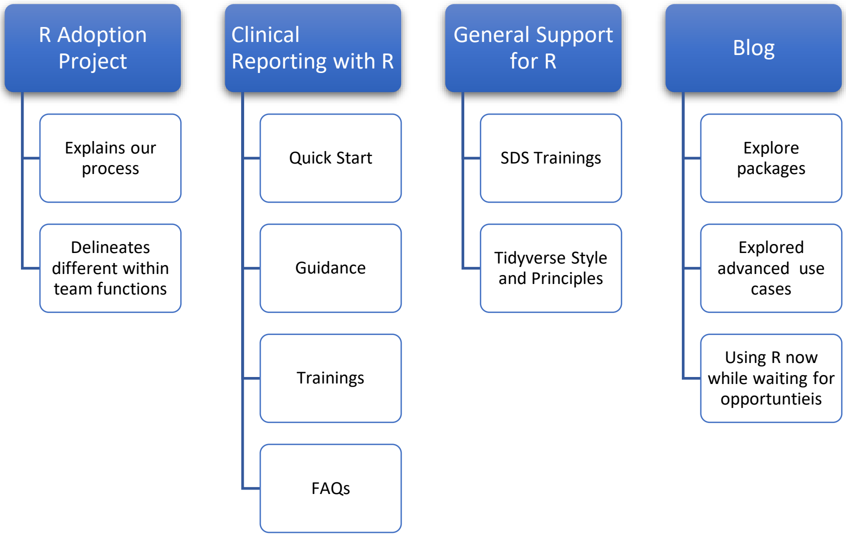
Q3 2021
- GSK-Atorus Release 2
- First R4QC SAC (TFLs)
- Target 40 studies R4QC Live

Q4 2021
- Target 5 R4QC SACs
- Project facing 100% trained in **tidyverse** and **ggplot2**

1H 2022
- R4Prod SAC

Figure 7: Homepage of the R for Clinical Reporting Website built using blogdown

The R for Clinical Reporting website serves as a clearinghouse for all things R and Clinical Reporting. As stated in the package name, blogdown, most of the websites that you can create through this package have a blog feature. I highly recommend the use of blogs, especially for those just starting their R journey. The blogs can serve as a multi-purpose tool. For example, Atorus (a CRO) has developed a package called Tplyr, which enables easy display creation. We had a blog post that discussed these exciting features from their package. At the same time, we have posts exploring the best ways to use more advanced R functions for Stage 2 QC of displays. However, the guiding principle for our website is everything needs to be in the context of clinical reporting. The website we have developed internally has 4 large blocks.



R Adoption Project
- Explains our process
- Delineates different within team functions

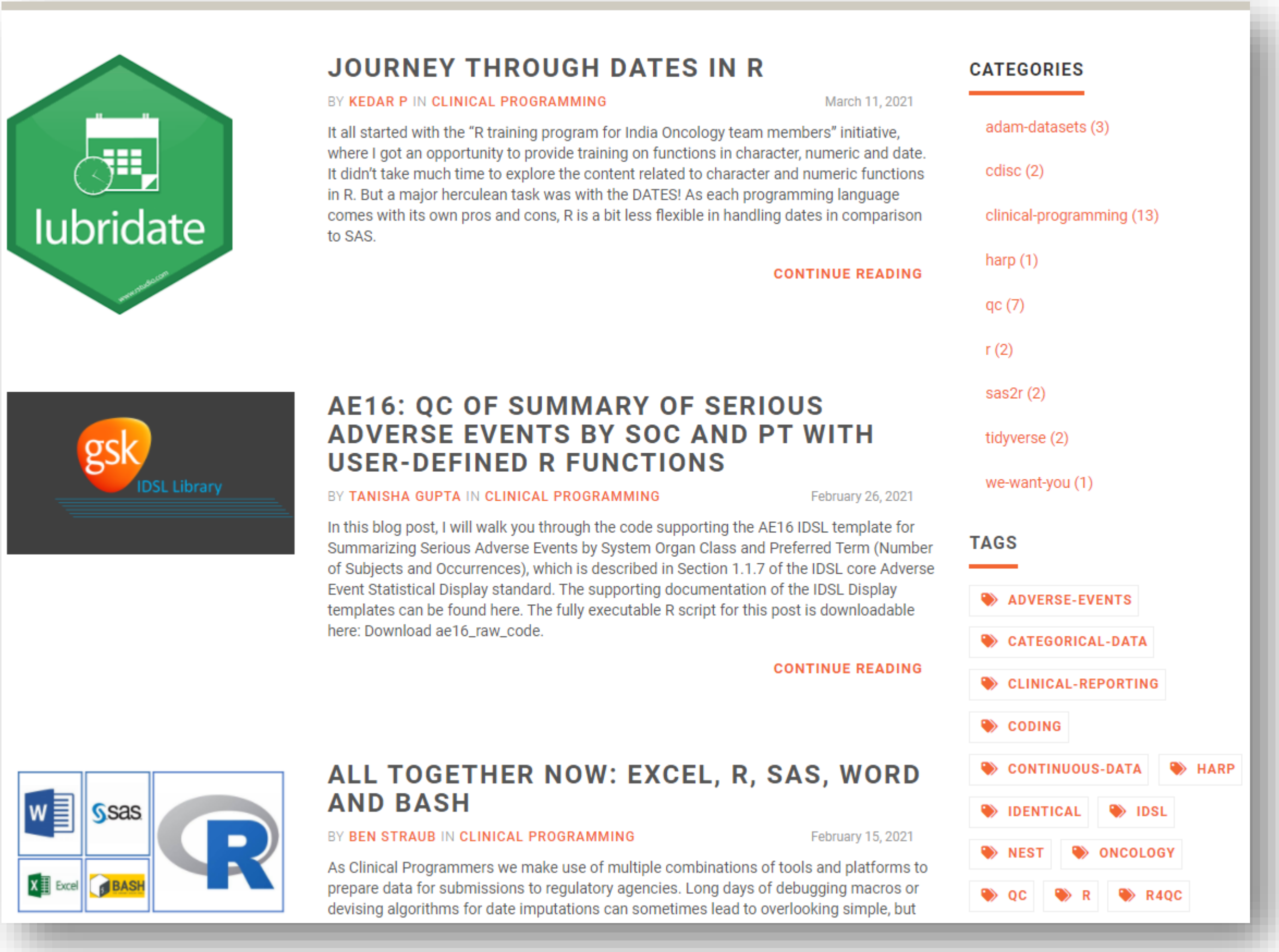
Clinical Reporting with R
- Quick Start
- Guidance
- Trainings
- FAQs

General Support for R
- SDS Trainings
- Tidyverse Style and Principles

Blog
- Explore packages
- Explored advanced use cases
- Using R now while waiting for opportuntieis

Figure 8: High-Level overview of Website Contents

Another positive side-effect of the website is the opportunities for programmers through blogging. A blog post is written in RMarkdown. Programmers must use their budding R skills to create a blog post. Creating the code is one thing but making it consumable to a public audience can take some extra time. This also serves as stopgap, that is while programmers are waiting to utilize R in their studies, they can also create blog posts using R. At GSK-CP we encourage programmers to look at packages, explore coding for specific display tables or using R for exotic situations in our programming ecosystem. The blog feature also allows for tagging and creating categories as seen below.



JOURNEY THROUGH DATES IN R

BY KEDAR P IN CLINICAL PROGRAMMING March 11, 2021

It all started with the "R training program for India Oncology team members" initiative, where I got an opportunity to provide training on functions in character, numeric and date. It didn't take much time to explore the content related to character and numeric functions in R. But a major herculean task was with the DATES! As each programming language comes with its own pros and cons, R is a bit less flexible in handling dates in comparison to SAS.

CONTINUE READING

AE16: QC OF SUMMARY OF SERIOUS ADVERSE EVENTS BY SOC AND PT WITH USER-DEFINED R FUNCTIONS

BY TANISHA GUPTA IN CLINICAL PROGRAMMING February 26, 2021

In this blog post, I will walk you through the code supporting the AE16 IDSL template for Summarizing Serious Adverse Events by System Organ Class and Preferred Term (Number of Subjects and Occurrences), which is described in Section 1.1.7 of the IDSL core Adverse Event Statistical Display standard. The supporting documentation of the IDSL Display templates can be found here. The fully executable R script for this post is downloadable here: Download ae16_raw_code.

CONTINUE READING

ALL TOGETHER NOW: EXCEL, R, SAS, WORD AND BASH

BY BEN STRAUB IN CLINICAL PROGRAMMING February 15, 2021

As Clinical Programmers we make use of multiple combinations of tools and platforms to prepare data for submissions to regulatory agencies. Long days of debugging macros or devising algorithms for date imputations can sometimes lead to overlooking simple, but

CATEGORIES

adam-datasets (3)
cdisc (2)
clinical-programming (13)
harp (1)
qc (7)
r (2)
sas2r (2)
tidyverse (2)
we-want-you (1)

TAGS

ADVERSE-EVENTS
CATEGORICAL-DATA
CLINICAL-REPORTING
CODING
CONTINUOUS-DATA
HARP
IDENTICAL
IDSL
NEST
ONCOLOGY
QC
R
R4QC

Figure 9: A snapshot of our latest blogs on the R for Clinical Reporting Website

CONCLUSION

The use of R within Clinical Programming at GSK is increasing at a rapid pace. The ability to share code and ideas through the packages blogdown and blogdown, in an elegant and visually appealing way, has aided our efforts immensely at R Adoption.   I would be remiss if I did not also espouse the virtues of Github, which allowed us to quickly copy code from the GSK-Statistical Data Sciences (SDS) and tailor to our needs.   Github is a powerful force within the open-source ecosystem and its ability to interact with RStudio enhances your possibilities to build great tools.

RECCOMENDED READING

- **BLOGDOWN:** https://bookdown.org/yihui/blogdown/
- **BOOKDOWN:** https://bookdown.org/yihui/bookdown

CONTACT INFORMATION

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

Ben Straub
GlaxoSmithKline
1250 S. Collegeville Road
Collegeville, Pennsylvania, US, 19426-0989
Email: ben.x.straub@gsk.com

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