

What R We Counting? *The Quest to Quantify R Usage in Biostatistics Organizations*

Benjamin Arancibia, GSK, Philadelphia, USA

ABSTRACT

As biostatistics organizations increasingly embrace open-source technologies, a pivotal question emerges: “How much R or Python are our clinical study teams actually using?” Though it’s a simple question, it’s incredibly complex when you begin to dive deep into it. This talk will explore how to tackle this challenge effectively. We’ll reveal practical methods to access and analyze usage data, craft actionable metrics, and engage stakeholders with compelling insights. Discover the challenges we faced in defining metrics within large enterprises and how we leveraged the GitHub API to gain a clearer picture of R adoption beyond the basics. Join us to learn how to turn data into actionable intelligence and elevate your understanding of open-source technology utilization in clinical study teams.

INTRODUCTION

The pharmaceutical industry is going through a rapid transformation as open-source technologies are being embraced for GxP processes. For biostatistics organizations, this has been the use of R as a primary programming language. As part of this rapid transformation and change, senior stakeholders who are accountable for objectives and metrics will come and ask – Hey! How much R and or Python are we using in our clinical trials? Though it is a relatively simple question, it can be difficult to answer it accurately with metrics that impact stakeholders. This paper provides an approach to answering this simple question with a focus on R as the primary language being measured.

Background Context

GSK Biostatistics has been on a journey to use more open source since 2017, but in 2020 we began together all the pieces. This includes platform, process, tools, pipeline, and training. In August 2023, the Biostatistics Leadership Team made two commitments: 1) all central tools must be built using open-source languages and 2) 50% of our code to be open-source by end of 2025. For more information on this journey, please watch our talk with Posit (Posit, 2024). With these commitments there has been greater scrutiny on what is the adoption of R in our clinical trials.

When a senior stakeholder in GSK Biostatistics asked the question, “How much R are we using on our clinical trials”, we began look at the current way of answering that question which was to manually enter and count the total number of outputs created by R. As part of our journey, we began to different pieces which includes platform, tools, pipeline, process, and training. One of the tools implemented is GitHub as GSK Biostatistics made the decision to move away from file shares and use version control as our way of working and collaboration. GitHub has a series of APIs¹ that allow users to interact with GitHub and create tools to better utilize its functionality. While working with the GitHub APIs to create tools to automate work for our study teams, we noticed that you could pull data from GitHub and with the transition of using GitHub as our version control, realized we could information about our study repositories to begin to answer the question “How much R are we using on our clinical trials”.

GitHub APIs

GitHub APIs allow for users to interact with repositories and pull data. When we were looking for an easy way to interact with the GitHub API, we began by looking for open-source packages. Ask, and the R Ecosystem provides! The {gh} package is a minimalist package to access GitHub APIs. Typically, it’s used by R Package developers for setting up personal access tokens (PAT) and other activities related to GitHub repository management, but we found that you query any API that is available (Csárdi, 2025). Utilizing the GitHub API is quite easy, the setup (code below) is quite basic and requires just a PAT.

```
install.packages("gh")
library(gh)
my_repos <- gh("GET /users/{username}/repos", username = "YOUR USERNAME")
vapply(my_repos, "[", "", "name")
```

This is a simple approach to get all repos tied to an individual’s username. More complex queries are also doable as seen below.

¹ Application programming interfaces (APIs), which are software connections that allow applications to communicate with each other

```

#function to get repo level information for all repos using graphql
get_all_repos<-function(auth = gh::gh_token()){

# 100 repos at a time - need to paginate thru
#(note that gh package offers limit=.Inf for REST API which paginates under the hood)
# also note there isn't a way to get contributors- keep using REST API for this
hasNextPage <- TRUE
endCursor <- NA
repo_list <- vector(mode = "list")

system.time(
  while(hasNextPage){

    query <- paste0('query {
      organization (login: ORGANIZATION NAME) {
        repositories (first: 100,
          after: ', jsonlite::toJSON(endCursor, auto_unbox = TRUE), ') {
          pageInfo {
            hasNextPage
            endCursor
          }
          edges{
            node{
              name
              diskUsage
              createdAt
              pushedAt
              updatedAt
              templateRepository{
                name
              }
              # Fetch only top 10 languages
              languages (first:10){
                totalSize
                edges {
                  size
                  node{
                    name
                  }
                }
              }
            }
          }
        }
      }
    }')
  }
}

```

By being able to access the GSK GitHub, we can pull data and begin to connect it with other data sources to get a better picture of what is the adoption rate of R.

Layers – What Matters to Stakeholders

As we began to look through the GitHub API and pull data to help answer the question “How much R are we using on our clinical trials” we encountered endless possibilities on what to count. Through the GitHub API, users can pull files counts, file size, commits, percentage of programming languages in repository, and other metrics. Edge cases also begin to appear which you will have to deal with, for example for repositories that are Shiny Apps, it is R code but when compiled the code is HTML and JavaScript. You will need to determine a way to count these repositories as part of R Adoption efforts.

Once determined what data, how to pull it, and how to make it align with the question you are trying to answer, engaging with stakeholders to understand what the metrics are they need to better understand R adoption is crucial.

Our Metrics

For GSK Biostatistics, engaging with senior stakeholders was crucial. To ensure we had an impact and deliver the information necessary, we asked what the most important topline metrics was they needed to understand our overall progress. The four metrics determined was:

- R Code size out of total code (R and SAS)
- Total Repos with any R usage
- Repos where R code size (out of R and SAS) is greater than Target Percent
- Growth of R: Rate of change of R code in a repo and organization

Where We Are Going

While creating ways to answer the question of the adoption of R, we found other ways to utilize the GitHub API to help support individuals in GSK Biostatistics.

One thing that we learned when rolling out GitHub and version control is that it is a difficult tool to learn. Version control is a different way to think about saving files compared to file shares. By looking at other pieces of data such as commit, changes during commits, number of branches, and merge conflicts – we can find repositories that need support. Proactively supporting studies provides a better experience and improved adoption of new tools.

Another area where we have explored is finding tools that are organically created by study teams. Finding these tools that fill specific needs allow us to help improve and support study teams. It also allows us to check if central tools we create are being used by study teams. Our goal is to support and improve the experience of study teams while adopting new tools.

CONCLUSION

Answering “How much R are we using on our clinical trials”, is tough when you begin to dig into all the layers that surround the question. Our use of the GitHub API is a way to do that. Each organization will have to define what it determines as success, but this programmatic approach allows for better metrics for senior stakeholders.

REFERENCES

Csárdi, G. (2025, Feb 24). *GH Package*. Retrieved from GitHub: <https://gh.r-lib.org/>
Posit. (2024, November). *GSK's R Journey: From Pilot Projects to Enterprise Adoption | Hosted by Posit*. Retrieved from <https://www.youtube.com/watch?v=xDrt6txplek>

CONTACT INFORMATION (HEADER 1)

(In case a reader wants to get in touch with you, please put your contact information at the end of the paper.)

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

Author Name: Benjamin Arancibia

Company: GSK

Email: Benjamin.c.arancibia@gsk.com

Website: <https://www.linkedin.com/in/bcarancibia/>