**Validating R for Pharma**

**Streamlining the Validation of Open-Source R Packages within Highly Regulated Pharmaceutical Work**

**Anuja Das, Juliane Manitz, Jaxon Abercrombie, Antal Martinecz, Doug Kelkhoff**

# Abstract

Using R in submissions to healthcare regulators often requires documentation showing that the quality of the programming packages used was adequately assessed. This can pose a challenge in R where many of the commonly used tools are open source. Through this paper, we will highlight the R Validation Hub's risk assessment framework for R packages that has been utilized by key pharma companies across the industry. We also showcase the products our working groups have developed including the {riskmetric} R package that evaluates the risk of an R package using a specified set of metrics and validation criteria, the {riskassessment} app that augments the utility of the {riskmetric} package within a Shiny app front end, and the repositories workstream that is working to develop a technical framework to maintain a 'repository' of R packages with accompanying evidence of their quality and the assessment criteria. Lastly, we will discuss the newly instituted Communications Workstream that seeks to bring individuals who are interesting in R Validation within the industry together in a variety of forums.

All our work is designed to facilitate the use of R within a highly regulated space and ease the burden of using R packages within a validated environment.

# Introduction

## The R Validation Hub

The R Validation Hub sits within the R Consortium and is a cross-industry collaboration to support the adoption of R within a biopharmaceutical regulatory setting through appropriate tools and resources that leverage the open source, collaborative nature of the language. It comprises a core set of workstream leads and team members along with participants from over 50 biotech and pharmaceutical companies who are brought together through regular community meetings, case studies, and other collaborations.

# The R Validation Hub was funded in 2018 and has since worked on developing several resources that are utilized industry-wide. While the initial impetus was for the group to consolidate definitions, regulations, and best practices for qualification and validation of the open-source language, the initial success has now given rise to a variety of workstreams that have worked on and are working on white papers, frameworks and prototypes, applications, and forums for open discussion. We will discuss several of these completed and ongoing efforts throughout this paper.

# The R Validation Hub White Paper

## A Reference for a Risk-Based Approach to R Validation

The white paper titled *A Risk-Based Approach For Assessing R Package Accuracy Within A Validated Infrastructure* was published in January, 2020. Authors of the paper represented several companies throughout the industry, however it is crucial to note that the framework presented reflects the thoughts of the R Validation Hub working group and not the companies themselves. Since its publication, several pharmaceutical companies have referenced this framework within their teams as they turn to R as a tool for regulatory submission.

The paper focuses on systems validation, not infrastructure validation, and highlights the need to test for accuracy, reproducibility, and traceability of R packages for intended use as part of the validation process, especially within the clinical submission space. It is also important to note that contrary to popular belief, the FDA does not mandate the use of any specific software for statistical analyses, per their Statistical Software Clarifying Statement.

## Base R Core Packages vs Contributed Open-Source Packages

The R Validation Hub makes a distinction between Base R packages and Open-Source Packages when assessing the risk of each.

The R Foundation develops the base and recommended packages and ensures a stringent SDLC process is followed for maintenance. Additionally, the R user community continually tests these packages and can provide feedback as needed. The various practices put in place to ensure validity of this set of packages automatically moves them towards a lower risk classification.

Open-Source packages can be developed by anyone who uses R and they do not typically go through a series of checks prior to going live. In this paper, the R Validation Hub has narrowed the parameters of contributed packages to those on CRAN since packages available on CRAN must pass a series of technical checks. A risk assessment is still necessary for validation of such packages.

## The Risk Assessment Framework

The recommended risk-based approach assesses Intended-For-Use packages based on Purpose, Maintenance Good Practice, Community Usage, and Testing. An overview of the proposed process is in the figure below:

A screenshot of a computer

Description automatically generated

In practice, this framework can be used to classify packages into overall categories of riskiness and consequent qualification or validation can be implemented accordingly. It is important to remember that following initial risk assessment and validation, long-term maintenance practices must be put in place so that the packages do not fluctuate into higher risk categories when later versions are released.

The Risk Metric package and Risk Assessment app discussed later in this paper are tangible products that were developed using this framework.

# R Package {riskmetric}

## Intended Use

The {riskmetric} package developed by the R Validation Hub encourages adoption of open-source packages by providing a tangible metric by which to evaluate package risk. Moving from perceived risk to quantified risk allows companies to evaluate the requirements more easily for validation, thereby leading to a more efficient timeline and method for adoption of packages into their infrastructure.

The package is helpful for various end users, from programmers to system administrators. It is intended to provide a foundation for companies to build their own risk assessment framework.

## How It Works

{riskmetric} works in three primary steps to assess the risk of an R package:

1. Finding a source of package information to compile package metadata
2. Assessing a package based on outlined validation criteria
3. Assigning a numeric risk score based on the results of risk criteria

A diagram of a computer code

Description automatically generated with medium confidence

The results of running a package through {riskmetric} are assembled in a dataset of validation criteria along with the resulting risk score.

## Future Work

The {riskmetric} is being continuously improved based on feedback, scenario planning, and complexities including, but not limited to, dependency footprints and package interoperability. Additionally, the push to make {riskmetric} more usable for non-technical or novice R users to increase the potential pool of end users resulted in the creation of the {riskassessment} application.

# The {riskassessment} Application

## Intended Use

The {riskassessment} app is a shiny front end that extends the principles of {riskmetric} to be more usable. The development of the shiny app essentially negates the need for technical coding prowess to reap the benefits of the risk assessment framework. Additionally, the app allows users to log assessments, generate reports, and store communication for assessed packages in one single repository.

## How it Works

While there is a demo application that is open to all and allows users to familiarize themselves with the interface, the actual {riskassessment} app has several pre-defined roles and requires authentication. You can use the app as an admin, lead, reviewer, or viewer.

Upon deployment, users are presented with the main screen where they can select packages and versions to assess. Once a package is selected, an overall risk classification is assigned, and users have the ability to check the results of each criterion using the Package Metrics tab. In the figure below, the {markdown} package was assessed and given a risk score of 0.19 with a Low-Risk classification.

A screenshot of a computer

Description automatically generated

To provide the overall context for both the risk assessment framework used and the resulting metrics, the application houses options to review assessment criteria, and build a report for the assessed package and resultant score. Additional capabilities are provided based on assigned roles. For instance, reviewers, leads, and admins can add comments to reports.

## Recognition

The {riskassessment} application continues to evolve with feedback and as the {riskassessment} metrics are adjusted. During the 2nd Shiny Conference in March 2023, the app was awarded the Best App by popular vote among 20+ other app submissions.

# Regulatory R Package Repository

## Intended Use

The Regulatory Repository working group is in its early stages and was initiated to engage in discussions on validation open-source software in regulated industries, and to simplify the burden of assembling and storing R packages with accompanying evidence and documentation of quality testing and validation assessments. The intention is to create a cross-industry repository to magnify the benefits of true open-source collaborations and ultimately de-risk the use of public R packages for regulatory submissions. The pilot will involve the development of a centralized and public repository along with its accompanying validation pipeline and reference container image.

## Questions Answered

The resulting framework from this workstream seeks to provide answers to questions that typically come up during the validation process:

1. What are the main differences in validation considerations between SAS and open-source languages?
2. What documentation and testing must be completed to prove the reliability and consistency of packages?
3. What existing resources and guidelines can be leveraged by companies as they develop their validation strategies?

Outputs from the workstream could also incorporate best practices that help teams prepare for audits when using open-source packages as part of their regulatory work, and that help build industry consensus on avoiding uncertainty in health authority response to submissions.

## Future Work

The pilot is currently ongoing with a slated readout in Q3 with an interim presentation at the user! Conference in July. The team will also be presenting at the Programming Heads Council in June.

The pilot will include a base reference image, basic automated validated pipeline, and a basic repository to store and manage validated R packages.

# The Communications Workstream

## Intended Use

The newest working group within the R Validation Hub is the Communications Workstream which seeks to engage and leverage the essence of all open-source work – the users. Engagement with our entire gamut of stakeholders, from data scientists within the industry to casual R users, ensures that the R Validation Hub is working on the right scope of work that brings the highest value to the open-source community within the pharmaceutical space.

## Forums

The Communications Workstream operates through four primary avenues:

1. The Website
2. Community Meetings
3. Industry Case Studies
4. Social Media + Company Presence

The Community Meetings cover a variety of topics that range from open discussions on future work, a focus on specific company case studies, and breakouts to talk about what validation means (the answer varies depending on who you talk to!). They have been a great way to ensure that R Validation Hub projects have been on the pulse of the most pressing issues at hand and are also a great way to generate feedback on ongoing efforts and use cases.

The Website functions as a repository for all the resources developed by the R Validation Hub, blog posts, and industry case studies that showcase examples of the R Validation Framework in action. You can also find out about all ongoing work and can sign up to get involved in whatever capacity you choose.

## Future Work

The working group is working to broaden the impact of the Validation Hub through a series of goals.

1. A website refresh that allows increased interaction with resources as well as providing updates on ongoing and future work
2. Updates to previous case studies to bring learnings and further use cases to the broader audience
3. Establishing connections with company R User Groups to identify commonalities within the pharmaceutical industry, and to identify areas that could lead to the creation of future working groups

# Conclusion

We are always looking for new volunteers and members. Please reach out to us if you want to get involved, or have any questions about our work.