External R Package Qualification Implementation at Merck: Case Study using GGally

Uday Preetham Palukuru, Pawel Bernecki, Jane Liao, and Yilong Zhang, Merck & Co., Inc., Kenilworth, NJ, USA

Introduction

There has been a growing interest in pharmaceutical industry to use R for clinical trial data analysis and reporting (A&R). Using R for regulatory submission purposes requires careful qualification of R packages given that the open-source packages differ in their quality of development. Many cross-industry initiatives including R Validation Hub and TransCelerate have published framework for qualifying R packages to be used in a regulatory setting (Nicholls, Bargo, & Sims, 2020) (Amoruccio, Lee, & Woodie, 2021). Our organization has been exploring the use of R in a regulatory setting for the past few years. A framework has been developed internally for qualifying external R packages that incorporates elements from both R Validation Hub and TransCelerate framework. This framework is currently being used to qualify both internally developed and externally sourced R packages for use in clinical trial A&R.

In this document, we demonstrate this risk-based package qualification framework using the GGally R package. We provide the workflow as well as relevant details regarding the package qualification process used to qualify GGally as a moderate risk R package. We hope this inspires other organizations to use R in a regulatory setting as well as generate discussion to improve our existing framework.

Risk-Based Package Qualification Framework

The R package qualification framework deployed at Merck is based on validation as defined by FDA (The R Foundation for Statistical Computing c/o Institute for Statistics and Mathematics, 2021). The goal of the framework is to create documentation that contains qualification details of R package based on pre-specified criteria. The framework employs a risk-based strategy to qualify R packages based on the type of A&R deliverable being generated. The types of deliverables and their associated R package risk levels are shown in Table 1.

Type of Deliverables	Example	R Package Risk
External (electronic Common Technical Document (eCTD))	Clinical Study Report (CSR) and submission package Drug labeling Agency request	Low
External (non-eCTD) Internal (Outside Department)	Data monitoring committee Manuscript & publication (using clinical data) Internal committee review or presentation	Moderate or Low
Exploration/Within Department	Data exploration Data quality checks Exploratory analysis	Open, Moderate, or Low

Table 1. Examples of deliverables and respective risk categories of the R packages

The pre-specified criteria used to qualify a R package into the desired risk category are defined as:

- c1: Package is developed and maintained by a trusted vendor.
- c2: Package is user-facing with sufficient software development lifecycle (SDLC) evidence equivalent to internal SDLC requirement.
- c3: Package is not user-facing and all packages dependent on this R package are qualified.
- c4: Package is user-facing with additional internal work to complete necessary steps following internal SDLC requirement.
- c5: Package maintained by a trusted person or organization.

An R package can be qualified under low risk category if it meets any of the first four criteria i.e., c1-c4. For moderate risk category the R package needs to meet any of the five criteria i.e., c1-c5. Any R package used for exploratory purposes and not qualified under either low-risk or moderate-risk category is categorized as open-risk. Any other R package deployed by users from external sources such as Comprehensive R Archive Network (CRAN) or other repositories is automatically categorized as open risk.

R Package Qualification Workflow

Within our organization a shared baseline strategy recommended by RStudio is followed to manage a reproducible R environment (RStudio, 2020). The defining characteristic of shared baseline strategy is that R package availability is tied to R installations using site-wide libraries. The use of scheduled updates to the site-wide libraries allows all users to use the same installed packages, thereby creating a baseline environment to share and re-run work. An R package within our organization is available through a regularly updated site-wide library installation called Global R library. The global R library is a set of directories containing installed R packages and their dependencies. There are 3 risk levels within the global R library corresponding to the risk category used in package qualification. The global R library is nested and independent, with all low-risk packages included in the moderate-risk library, and all moderate-risk packages included in the open-risk library. Our organization employs RStudio Package Manager (RSPM) as the R package repository server to host source code to install the packages in global R library. A high-level global R library update and package qualification workflow is summarized below (Figure 1).

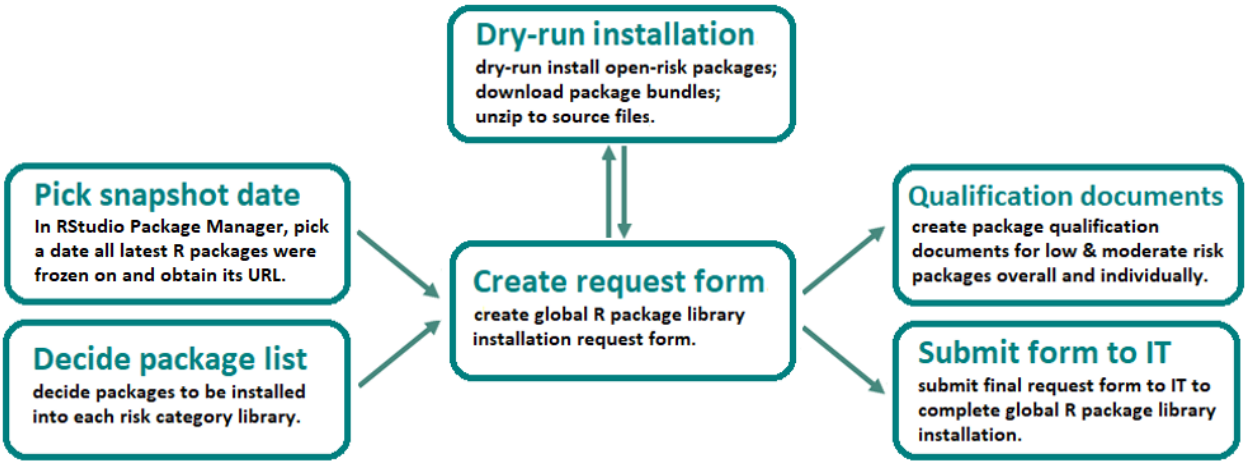


Figure 1. Global R library update and package qualification workflow diagram

As the process to qualify R packages can be cumbersome, an internally developed R package is used to streamline and automate the package qualification process. Once the qualification documents are generated, a panel comprising of qualified statistician and statistical programmer reviews the documentation for validity and accuracy.

GGally Package Qualification:

GGally is an R package that extends ggplot2 R package functionality by adding several functions to reduce the complexity of combining geometric objects (geoms) with transformed data (Schloerke, 2020). Based on a request to use the GGally package in a publication, the package was qualified under moderate risk category. The steps in the qualification process followed were:

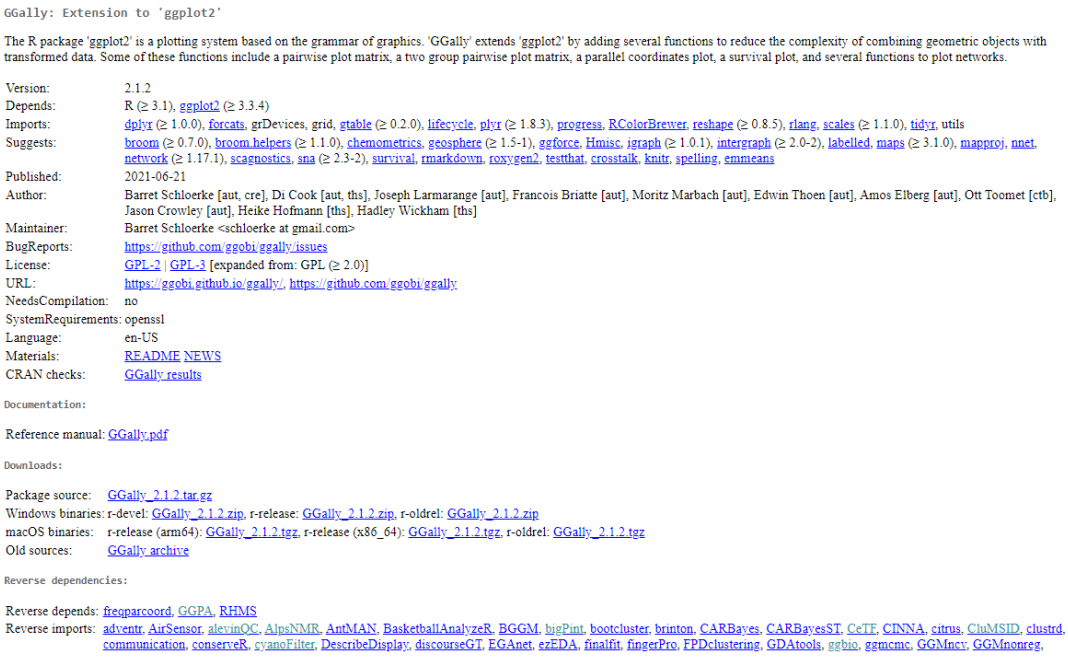
1. Review package documentation to determine qualifying criteria. It was determined that this package can be qualified using the c2 and c5 criteria.
2. Perform a dry-run installation for global R library update, with GGally set as moderate risk.
3. Check installation log for errors / warning messages (Figure 2):

```
[1] "trying URL 'https://rspm.merck.com/open_risk/2022-03-07+OToyMDk1MywxMDoyOTU4LGNyYW47OUVFQ0RBODA/src/contrib/GGally_2.1.2.tar.gz'"
[2] "Content type 'application/x-gzip' length 1432754 bytes (1.4 MB)"
[3] "=================================================="
[4] "downloaded 1.4 MB"
[5] "Read 2 items"
[6] "* installing *source* package ‘GGally’ ..."
[7] "** package ‘GGally’ successfully unpacked and MD5 sums checked"
[8] "** using staged installation"
[9] "** R"
[10] "** data"
[11] "*** moving datasets to lazyload DB"
[12] "** inst"
[13] "** byte-compile and prepare package for lazy loading"
[14] "** help"
[15] "*** installing help indices"
[16] "** building package indices"
[17] "** testing if installed package can be loaded from temporary location"
[18] "** testing if installed package can be loaded from final location"
[19] "** testing if installed package keeps a record of temporary installation path"
[20] "* DONE (GGally)"
```

Figure 2. GGally dry run installation log snippet.

4. Check package code coverage and associated SDLC documentation (criterion c2) (Figure 3, Figure 4):

Figure 3. Code coverage statistics of GGally obtained from internal RSPM server.



GGally: Extension to 'ggplot2'

The R package 'ggplot2' is a plotting system based on the grammar of graphics. 'GGally' extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data. Some of these functions include a pairwise plot matrix, a two group pairwise plot matrix, a parallel coordinates plot, a survival plot, and several functions to plot networks.

Version:	2.1.2
Depends:	R (≥ 3.1), ggplot2 (≥ 3.3.4)
Imports:	dplyr (≥ 1.0.0), forcats, grDevices, grid, gtable (≥ 0.2.0), lifecycle, plyr (≥ 1.8.3), progress, RColorBrewer, reshape (≥ 0.8.5), rlang, scales (≥ 1.1.0), tidyr, utils
Suggests:	broom (≥ 0.7.0), broom.helpers (≥ 1.1.0), chemometrics, geosphere (≥ 1.5-1), ggforce, Hmisc, igraph (≥ 1.0.1), intergraph (≥ 2.0-2), labelled, maps (≥ 3.1.0), mapproj, nnet, network (≥ 1.17.1), scagnostics, sna (≥ 2.3-2), survival, rmarkdown, roxygen2, testthat, crosstalk, knitr, spelling, emmeans
Published:	2021-06-21
Author:	Barret Schloerke [aut, cre], Di Cook [aut, ths], Joseph Larmarange [aut], Francois Briatte [aut], Moritz Marbach [aut], Edwin Thoen [aut], Amos Elberg [aut], Ott Toomet [ctb], Jason Crowley [aut], Heike Hofmann [ths], Hadley Wickham [ths]
Maintainer:	Barret Schloerke <schloerke at gmail.com>
BugReports:	https://github.com/ggobi/ggally/issues
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2.0)]
URL:	https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally
NeedsCompilation:	no
SystemRequirements:	openssl
Language:	en-US
Materials:	README NEWS
CRAN checks:	GGally results

Documentation:

Reference manual: GGally.pdf

Downloads:

Package source:	GGally_2.1.2.tar.gz
Windows binaries:	r-devel: GGally_2.1.2.zip, r-release: GGally_2.1.2.zip, r-oldrel: GGally_2.1.2.zip
macOS binaries:	r-release (arm64): GGally_2.1.2.tgz, r-release (x86_64): GGally_2.1.2.tgz, r-oldrel: GGally_2.1.2.tgz
Old sources:	GGally archive

Reverse dependencies:

Reverse depends:	freqparcoord, GGPA, RHMS
Reverse imports:	adventr, AirSensor, alevinQC, AlpsNMR, AntMAN, BasketballAnalyzeR, BGGM, bigPint, bootcluster, brinton, CARBayes, CARBayesST, CeTF, CINNA, citrus, CluMSID, clustrd, communication, conserveR, cyanoFilter, DescribeDisplay, discourseGT, EGAnet, ezEDA, finalfit, fingerPro, FPDclustering, GDAtools, ggbio, ggmcmc, GGMncv, GGMnonreg,

Figure 4. GGally SDLC documentation from CRAN.

5. Cross-check against internal database (White List) containing trusted package authors / organizations (criterion c5). R Package Author White List is a list of trusted R package authors (person or organization) identified by our organization's Subject Matter Experts (SME).
6. Run program using internally developed R package to generate the qualification document. The details included in the qualification document are as shown below:

Package Qualification – "GGally"

Qualification Overview

The purpose of this document is to demonstrate that GGally, when used in a qualified fashion, can support the appropriate regulatory requirements for validated systems, thus ensuring that resulting electronic records are "trustworthy, reliable and generally equivalent to paper records."

Package	GGally
Risk level	moderate
Qualification date	2022-03-14
Qualification criteria	c2, c5

Package Information

package	GGally

version	2.1.2
author	Barret Schloerke [aut, cre], Di Cook [aut, ths], Joseph Larmarange [aut], Francois Briatte [aut], Moritz Marbach [aut], Edwin Thoen [aut], Amos Elberg [aut], Ott Toomet [ctb], Jason Crowley [aut], Heike Hofmann [ths], Hadley Wickham [ths]
maintainer	Barret Schloerke <schloerke@gmail.com>;
license	GPL (>= 2.0)
description	The R package 'ggplot2' is a plotting system based on the grammar of graphics. 'GGally' extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data. Some of these functions include a pairwise plot matrix, a two group pairwise plot matrix, a parallel coordinates plot, a survival plot, and several functions to plot networks.
url	https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally
bugreports	https://github.com/ggobi/ggally/issues
systemrequirements	openssl

Qualification Details

Criteria C2

Criteria c2: There is sufficient evidence of publicly available software development lifecycle information, including authors, source code, test cases, release notes, and user guides.

To qualify GGally, we reviewed and confirmed the R package GGally follows a proper software development lifecycle.

- Each exported (user facing) function contains documentation.
- The released version has a unique version number on CRAN.
- The R package development use a version control system. https://github.com/ggobi/ggally
- Proper testing has been provided with source code. The code coverage is 86%.
- The R package passed a series compliance check through CRAN.
- The R package has a maintainer, Barret Schloerke schloerke@gmail.com
- The R package has a bug report approach using https://github.com/ggobi/ggally/issues
- The R package can be properly built and installed on system.
- Source code archive files ("tarballs") are made available via the CRAN mirror infrastructure.
- All current and historical released versions of this R package are available from the main CRAN server (http://cran.r-project.org/src/base/) and its worldwide mirrors (http://cran.r-project.org/mirrors.html).

Criteria C5

- Criteria c5: Package maintained by a trusted person or organization listed in R package author white list
- The released version has a unique version number on CRAN.
- The R package passed a series compliance check through CRAN.
- The R package has a maintainer, Barret Schloerke schloerke@gmail.com
- The R package can be properly built and installed on system.

- Source code archive files (“tarballs”) are made available via the CRAN mirror infrastructure.
- All current and historical released versions of this R package are available from the main CRAN server (http://cran.r-project.org/src/base/) and its worldwide mirrors (http://cran.r-project.org/mirrors.html).

7. After the qualification document is generated, panel compromising of a qualified statistician and a statistical programmer reviews the document for accuracy and validity.

After the qualification was completed, GGally was included in the moderate risk category update for global R library formal installation.

Conclusion

A risk-based R package qualification process has been deployed at Merck to classify R packages based on generated A&R deliverables. This process has been automated using internally developed R package to both streamline the process as well as reduce any human errors. The qualification of GGally R package under moderate risk category using the qualification process, demonstrates the useability of the qualification framework for qualifying R packages in a regulatory setting. There is ongoing work to enhance the internally developed R package used in package qualification framework to further automate the process. We are also working on defining the pre-specified criteria used to qualify an organization or group of vendors as trusted source, thereby reducing the burden of qualifying individual packages.

References

Amoruccio, V. J., Lee, M., & Woodie, D. (2021). A TransCelerate Initiative – How Can You Modernize Your Statistical Environment. PharmaSUG 2021. SI-028. PharmaSUG. Retrieved April 11, 2022, from https://www.pharmasug.org/proceedings/2021/SI/PharmaSUG-2021-SI-028.pdf

Nicholls, A., Bargo, P. R., & Sims, J. (2020, January 23). A risk-based approach for assessing R package accuracy within a validated infrastructure. Retrieved April 11, 2022, from https://www.pharmar.org: https://www.pharmar.org/white-paper/

RStudio. (2020). Shared Baselines. (RStudio) Retrieved April 11, 2022, from Reproducible Environments: https://environments.rstudio.com/shared

Schloerke, B. (2020, March 25). GGally: Extension to ggplot2. Retrieved April 11, 2022, from https://www.rdocumentation.org/packages/GGally/versions/1.5.0

The R Foundation for Statistical Computing c/o Institute for Statistics and Mathematics. (2021, October 18). R: Regulatory Compliance and Validation Issues, A Guidance Document for the Use of R in Regulated Clinical Trial Environments. Vienna, Austria. Retrieved April 11, 2022, from https://www.r-project.org/doc/R-FDA.pdf

Corresponding Author Contact

We encourage feedback to improve our framework and processes. For any questions or feedback please reach out to preetham.palukuru@merck.com