

R “Validation” Hub Meeting

02 April 2019

Attendees:

Andy Nicholls (GSK) – Meeting Chair
Lyn Taylor (Phastar) – Meeting Secretary

Alexander Lock-Achilleos (GSK)
Alun Bedding (Roche)
Anthony Williams (Fred Hutchinson CRC)
Bob Engle (Biogen)
Chris Toffis (Syne qua non) – PSI AIMS SIG
Doug Kelkhoff (Roche)
Dharmesh Desai (BioMarin Pharmaceutical)
Eric Milliman (Biogen)
Greg Cicconetti (Abbvie) – ASA BIOP Software
John Sims (Pfizer)
Joseph Rickert (R-Studio)
Juliane Manitz (Merck Serono)
Keaven Anderson (Merck) – ASA BIOP Software
Magnus Mengelbier (LimeLogic)
Markus Elze (Roche) – PSI AIMS SIG
Matthias Trampisch (Boehringer-Ingelheim)
Michael Carniello (Astellas)
Nate Mockler (Biogen)
Nicollo Bassani (Quanticate)
Patric Stracke (Sanofi)
Paul Schuette (FDA CDER)
Rinki Jajoo (Merck)
Satish Murthy (J&J)
Steve Noga (RHO Inc.)
Tilo Blenk (GSK)
Tomas Drgon (FDA)
Yilong Zhang (Merck)

Previous Action Items

Action Item	Assigned team member(s)	Deadline	Status
Communication, dissemination of information about the project to the wider group Possible Funding discussion if we need to write a package to pull metrics	Andy Nicholls/ Lyn Taylor/ Joe Rickert	May 2019	Ongoing
Put a package through metrics gathering process and develop prototype report	Andy	May 2019	Ongoing
Review of the website content: Direct review back to Andy or message via slack as you prefer.	Kieran Martin, Min Lee, Juliane Manitz, John Simms,	May 2019	Ongoing

R “Validation” Hub Meeting

02 April 2019

	Kieven Andersen, Alex lock-Achileos		
Risk assessment workstream – set objectives + milestones	Yilong Zhang , Rebecca Krouse, Doug Kelkhoff, Matthias Trampisch, Eric Nantz	28 th March	Closed
Requirements/tests workstream – appoint spokesperson and set objectives + milestones	Nate Mockler, Keaven Anderson, Tilo Blenk.	May 2019	Ongoing
Validation white paper workstream – appoint spokesperson and set objectives + milestones	Andy Nicholls, Paulo Bargo	May 2019	Ongoing
Present to the group at the next meeting on how R Shiny may be used to push the live metrics from GitHub to a website.	Matthias Trampisch	4 th March	Closed
Write text for the overview website page and request review by website review team	Min Lee (and then Kieran Martin, Juliane Manitz, John Simms, Kieven Andersen, Alex lock-Achileos to review)	May 2019	Ongoing
Update website to put the Validation as the main focus and bring minutes/presentation/ About us more together.	Andy	28 th March	Closed
Select appropriate packages to use as proof of concepts prior to getting additional resource for expanding the task	Yilong / risk assessment workstream	28 th March	Closed
Write blog post on gathering metrics and using these in a risk assessment	Doug Kelkhoff	TBC	Open
Investigate how to educate people in the structure of R packages. Yilong suggested including on the website an overview of the structure – so if someone has high level questions we can direct them to the information. Andy added this could be in the form of a FAQ page.	TBC	May 2019	Ongoing
Some members of the team found it difficult to access via Skype and preferred the Webex method. Lyn to set up a poll and to send with the meeting minutes to see how many people have an issue and if there is preference for Webex then we can try to move back to that method. Once poll returned, will set up next meeting.	Lyn	28 th March	Closed – andy webex set up
Andy wants to put the company logo's on the website rather than a list of the companies. All to contact their company to ask if this is permissible. Andy to ensure that the text on the website states that the listed companies participate in the project rather than any endorsement.	All	28 th March	Closed

R “Validation” Hub Meeting

02 April 2019

Meeting Agenda

- R Shiny App to collect metrics: Matthias Trampisch (Boehringer-Ingelheim)
- Streams Progress updates: Streams leaders

Discussion

R Shiny App to collect metrics

Matthias Trampisch (Boehringer Ingelheim) presented on the R Shiny App that he’s written to push the live metrics from GitHub/CRAN into a word document report. Some aspects of the report require manual entry however, where possible, data is scraped from the relevant sites to create the key content for the report. This report is used as a Qualification document to evaluate the risk associated with each package. Matthias keeps a list of the packages and versions that have been evaluated. The App currently just runs from the latest version of the package available.

Using ggplot2 as an example, Matthias shared with the team the Qualification document that the R Shiny App creates (see further detail below). The App collects data from CRAN, and also takes information directly from the package website to provide a comparison of the two sources.

At Boehringer, initially only 16 packages were directly requested to be qualified. However, due to the dependencies this led to 92 being evaluated. It is worth noting that some packages are wasteful with dependencies (i.e. Surrogate is very “wasteful” with as many as 146 dependencies). Matthias confirmed that the App finds all dependencies and not just those that are actually called by the code used.

Given the App uses web scraping code, one current limitation is that if the website changes, then the App has to be updated to be able to still collect the data. Joe Rickert suggested to use `package_dependencies` function and `packagefinder`. Matthias confirmed he’d used the `igraph` package which is useful to sort the dependencies to ensure that from top to bottom there are no missing dependencies. Joe suggested those packages could also be used to show dependencies between package authors. These groups of authors, or companies could contribute to the evidence for lower risk packages.

The App presented can be used to provide documentation of the risk associated with r packages used for clinical data analysis. However, currently, at Boehringer there is also structural testing of the functional call by reproduction in SAS or if it can’t be reproduced another option is to use expert opinion to verify the code is doing what we expect.

R “Validation” Hub Meeting

02 April 2019

The qualification document presented includes the following sections:

- 1) Package summary which is scraped from CRAN

1.1 Package summary

- **Package:** `ggplot2` (<http://bi-cranmirror//web/packages/ggplot2>)
- **Origin:** CRAN
- **Author:** Hadley Wickham [[aut.](#), [cre.](#)], Winston Chang [[aut.](#)], RStudio [[cph.](#)]
- **Maintainer:** Hadley Wickham ([hadley at rstudio.com](mailto:hadley@rstudio.com))
- **Depends on 34 packages:** `ggplot2`, `digest`, `grid`, `gtable`, `MASS`, `plyr`, `reshape2`, `scales`, `stats`, `tibble`, `lazyeval`, `Rcpp`, `stringr`, `RColorBrewer`, `dichromat`, `munsell`, `labeling`, `R6`, `viridisLite`, `cli`, `crayon`, `methods`, `pillar`, `rlang`, `utils`, `glue`, `magrittr`, `stringi`, `colorspace`, `assertthat`, `grDevices`, `utf8`, `tools`, `graphics`

- 2) Package Information

1.2 Package information, directly taken from site

Version: 2.2.1

Depends: R (= 3.1)

Imports: `digest`, `grid`, `gtable` (= 0.1.1), `MASS`, `plyr` (= 1.7.1), `reshape2`, `scales` (= 0.4.1), `stats`, `tibble`, `lazyeval`

Suggests: `covr`, `ggplot2movies`, `hexbin`, `Hmisc`, `lattice`, `mapproj`, `maps`, `maptools`, `mgcv`, `multcomp`, `nlme`, `testthat` (= 0.11.0), `quantreg`, `knitr`, `rpart`, `rmarkdown`, `svglite`

- 3) Package short summary

1.3 Package short summary

A system for ‘declaratively’ creating graphics, based on “The Grammar of Graphics”. You provide the data, tell ‘ggplot2’ how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

- 4) Package history: To assess how well the package has been maintained, the app counts the number of previous versions released to CRAN by looking at the archived versions of the package. Also, within the source of the package, you can see if the package has testing routines, located in package name / tests folder. The App counts the functions under this folder and reports the number of tests the author has provided for the package. The download statistics are also output.

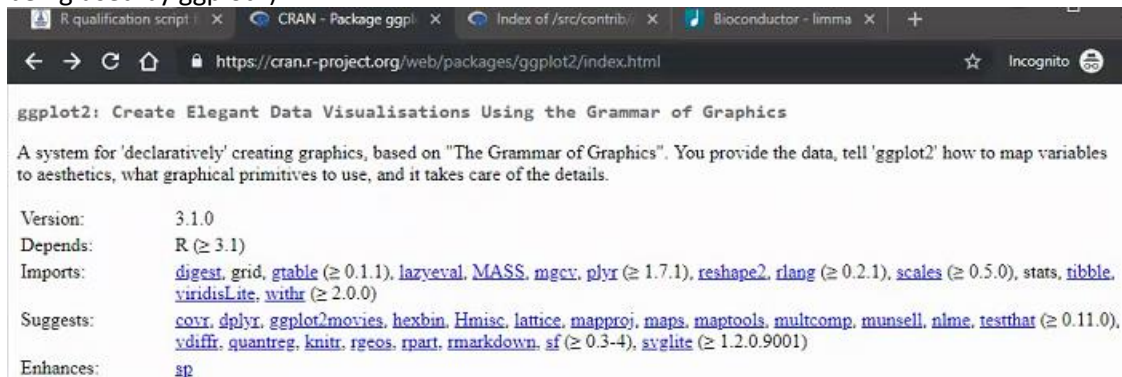
1.4 Package history

This CRAN package has been updated 32 times. The first version was submitted as `ggplot2_0.5.1.tar.gz` (10-Jun-2007 21:49) and the last version before the current one as `ggplot2_3.0.0.tar.gz` (03-Jul-2018 19:20). The author provides a total of 236 testing programs with the package. The package has been downloaded a total of 19,540,424 times from the Rstudio servers since the beginning of cranlogs in October 2012. The package has been downloaded 4,336,025 times in the last 180 days (between 2018-10-04 to 2019-04-02).

R “Validation” Hub Meeting

02 April 2019

- 5) Requirement check for dependencies/imports: This section lists the dependent packages and if they are from Base, Recommended, CRAN or Bioconductor sources. In order for this package (ggplot2) to pass qualification, all the packages it is dependent on must have already passed the process. Any packages missing from this qualification process are listed so that they can be assessed prior to continuing to assess this package. Note that the Imports listed on CRAN may not be exhaustive (i.e. 14 shown below, but this requirement check found 40 packages were being used by ggplot2).



The screenshot shows the CRAN page for the ggplot2 package. The browser address bar indicates the URL is https://cran.r-project.org/web/packages/ggplot2/index.html. The page title is "ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics". The description states: "A system for 'declaratively' creating graphics, based on 'The Grammar of Graphics'. You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details." The page lists the following information:

- Version: 3.1.0
- Depends: R (≥ 3.1)
- Imports: digest, grid, gtable (≥ 0.1.1), lazyeval, MASS, mgcv, plyr (≥ 1.7.1), reshape2, rlang (≥ 0.2.1), scales (≥ 0.5.0), stats, ribbon, viridisLite, withr (≥ 2.0.0)
- Suggests: covr, dplyr, ggplot2movies, hexbin, Hmisc, lattice, mapproj, maps, maptools, multcomp, munsell, nlme, testthat (≥ 0.11.0), ydiff, quantreg, knitr, rgeos, rpart, rmarkdown, sf (≥ 0.3-4), svglite (≥ 1.2.0.9001)
- Enhances: sp

2 Requirement check (check if all imported packaged are qualified)

The Requestor checks if required packages (imports) are already qualified and available in the Qualified R Library on the CRE. If a package fails to be available, that package has to be qualified first.

2.1 Available packages and version

There are 40 of the 40 dependent packages available.

Package	base_R	recommended_R	CRAN	Bioconductor	Version
assertthat	-	-	x	-	0.2.0
cli	-	-	x	-	1.0.1
colorspace	-	-	x	-	1.3-2
crayon	-	-	x	-	1.3.4

- 6) Acceptability and Applicability (including Author and licensing): This section is mostly manual and can copied in from the package itself or internet. The section does include automatically any citations from the author if listed on CRAN with the package. For example, the text below is automatically created by the App, however can be easily expanded as shown.

3 Acceptability and applicability

The Requestor should state basic information about the author and the package available in books, publications and/or articles. This should include the information from the citation information given on CRAN (or use the R command citation("ggplot2")) and a list of reverse import(s) of package. State if any related R packages or SAS procedures are known to exist (to the best knowledge).

Verify that the license requirements of the package only include General Public Licenses (GPL). If any other license is required, the package fails the qualification.

R “Validation” Hub Meeting

02 April 2019

3.1 Author (state basic information about author of package, if available)
(any available information about author)

3.2 Books, publications and/or articles about the package (list at least citation information from CRAN)

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
(any additional information)

3.3 Related R packages or SAS procedures
(any additional information)

3.1 Author (state basic information about author of package, if available)

Hadley Wickham

- Bachelor of Human Biology and a B.Sc. and M.Sc. in statistics from the University of Auckland
- PhD at Iowa State University
- currently Chief Scientist at [RStudio](#) and adjunct professor at Stanford University, University of Auckland and Rice University
- develops open source R tools for data science, data import and software engineering
- was awarded the John Chambers Award for Statistical Computing in 2006 and called Fellow of the American Statistical Association in 2015

Winston Chang

- Ph.D. in Psychology from Northwestern University
- created a website called "Cookbook for R" during his time as a graduate student, which contains recipes for handling common tasks in R
- currently software engineer at [RStudio](#), where he works on data visualization and software development tools for R
- author of *R Graphics Cookbook*

[RStudio](#) is a copyright holder.

3.2 Books, publications and/or articles about the package (list at least citation information from CRAN)

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

Relevant publications and books (only a few mentioned here):

- Ito, K., & Murphy, D. (2013). Application of *ggplot2* to [pharmacometric](#) graphics. *CPT: pharmacometrics & systems pharmacology*, 2(10), 1-16.
- Muenchen, R. A. (2011). Graphics with *ggplot2*. In *R for SAS and SPSS Users* (pp. 521-598). Springer, New York, NY.
- Kahle, D., & Wickham, H. (2013). *ggmap: Spatial Visualization with ggplot2*. *R Journal*, 5(1).
- Wollschläger, D. (2017). Diagramme mit *ggplot2*. In *Grundlagen der Datenanalyse mit R* (pp. 607-628). Springer Spektrum, Berlin, Heidelberg.
- Tyner, S., Briatte, F., & Hofmann, H. (2017). Network Visualization with *ggplot2*. *The R Journal*.
- Teutonico, D. (2015). *ggplot2 Essentials*. Packt Publishing Ltd.
- Torgo, L. (2016). *Data mining with R: learning with case studies*. CRC press.
- Moon, K. W. (2016). *Learn ggplot2 Using Shiny App*. Springer, Cham.

3.3 Related R packages or SAS procedures

There are several quite different R packages providing tools for making graphics, but *ggplot2* is one of the most widely used ones. It is a plotting system for R, based on the grammar of graphics. There may be some overlaps with other packages, e.g. the *lattice* package. But

7) Licenses: There are currently 30 licenses on CRAN, the lawyers at Boehringer looked at some packages which actually didn't specify that commercial use was allowed by the license. They didn't say it was disallowed, but also didn't state that it was allowed. Therefore, this also needs to be checked when assessing a package.

Joe Rickert agreed that licensing is also really important. The R consortium Core Infrastructure Initiative (CII) group is also interested in the licensing issue and they may be helpful and willing to work with our group. Mark Hornick has also blogged on the topic (<https://blogs.oracle.com/author/mark-hornick>).

R “Validation” Hub Meeting

02 April 2019

3.4 Licenses

The package has the following license(s): GPL-2 | file LICENSE
These licenses do not impose any restrictions to commercial use.

The following licenses restrict commercial use, please contact the R Library Manager for further instructions:

- ACM Software License Agreement
- Creative Commons Attribution-NonCommercial 3.0 Unported License
- Creative Commons Attribution-NonCommercial-NoDerivs 3.0 United States License
- Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License
- Creative Commons Attribution-NonCommercial 4.0 International License
- Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License
- Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License

(any additional information)

- 8) Reverse Dependences (how often this package is used in other packages): The App lists and counts the number of reverse dependencies.

4 Reverse depends and reverse imports

Reverse depends and reverse imports are packages which require the qualified package to be functional. If many packages are listed here, this may be considered a surrogate for quality and safety of the proposed package.

4.1 Reverse depends

This package has 299 reverse dependencies; the following packages depend upon it. *ACSNMineR, afmToolkit, alakazam, AmpliconDuo, aoristic, apsimr, BatchMap, bayesDP, BCellMA, bde, biomod2, bootnet, braidReports, brms, btergm, CA3variants, caret, CAvariants, centralplot, ceterisParibus, cjoint, classifierplots, climwin, clustrd, CNVScope, coefplot, corkscrew, cowplot, cr17, crmPack, Crossover, CRTgeeDR, cystiSim, cytofan, dae, dataMaid, Deducer, DendroSync, DengueRT, DepthProc, dfexplore, dggridR, diffeR, difNLR, DoTC, dotwhisker, dslice, DTRlearn, dtwSat, dynOmics, dynr, earlywarnings, eeptools, egg, eiCompare, EnsCat, EpiCurve, ESGtoolkit, EventStudy, FactoClass, factoextra, fbroc, fishmove, flippant, flowDiv, forestmodel, fpp2, freqparcoord, frequency, gapfill, gapmap, gcerisk, GENASphere, genlogis, genomeplot, geomnet, geotoolsR, ggallin, ggalluvial,*

- 9) Additional information

5 Additional information given by the Requestor

The Requestor should add the following information:

- *Proposed application of the R package within BDS analyses (scope). If the R package is being qualified because it is a dependency for a more sophisticated package this may also be given as justification.*
- *(optional) working example of proposed usage*

(any remarks about package go here)

- 10) Acceptance criteria Table: summarizes if the package passes qualification or not.

Criteria	Comment	From
Acceptability and applicability	See package summary and the reverse import section (additional entries)	Requestor
Package Maturity	State information about package history, etc (additional entries)	Requestor
Justification for inclusion	State information about advantages, etc (additional entries)	Requestor

R “Validation” Hub Meeting

02 April 2019

Risk Assessment steam update: Metrics Automation

Yilong Zhang provided an update on the streams progress: The team have created a GitHub repo on the pharmaR page and put together the following list of Goals, Scope and achievements.

GitHub, Inc. [US] | <https://github.com/pharmaR/riskmetric/blob/master/reports/riskmetric.md>

Proof of Concept for Risk Metrics

Goal

- Overall Goal: An automatically pipline to create metrics for R packages on CRAN/Bioconductor

Scope

- Create a database that contain the risk metrics for every R package of each version.
- Define risk metrics
- Derive code coverage

Out-of-Scope

- Define the risk score based on risk metrics (pending confirmation)

Achievement as of 03/31/2019

- Create a [function to collected risk metrics](#) for five R packages as pilot
- Reviewed the R in Pharma validation page and proposed additional [suggested metrics](#)

Ongoing Items

- [Code coverage derivation](#)
- [Define number of downloads](#)

Reference

- [R in Pharma Validation](#) (link may broke as the page is under development)
- [R studio Package Selection](#)
- [Quantifying R Package Dependency Risk](#)

The streams aim is to maintain a public database similar to what Matthias showed but we still need to define the risk metrics and work out how to derive the code coverage. Code coverage is problematic as we'd need all packages to be run through and it would be very time consuming.

Further discussion is needed regarding how we could potentially calculate or given guidance on a summary risk score. This would be difficult as each company may have different levels of risk. However, as a group, we need to provide some guidance on it to help people interpret the metrics. It's our duty to make a recommendation to our colleagues.

For example, if our risk score is from 0-100, then it would be up to the company to decide what score is an OK to them to feel confident in using a package. Once useful source could be to look at literature on engineering “Risk associated with type of failure”.

R “Validation” Hub Meeting

02 April 2019

Magnus Mengelbier highlighted that there is a lot of discussion on how good code coverage is an indicator on risk. He identified another issue is that simply counting the tests written on a package may not be enough since we also need to check the quality of the tests.

The team may have to work on the idea that when we are looking at overall score based on many metrics, it's an initial score. If a package receives a high risk score then it's telling us to explore further into that package and in those cases we'd look at tests and code coverage in more detail. However, many packages are known to be low risk and hence if they get a very low score it's highlighting that we don't need to look any deeper. So our role as a group may be more about us communicating how to interpret the risk to give people confidence of when and when not to use R in Pharma.

For Matthias's qualification app, he requires 3 criteria to be positive in order to pass the package for use. Matthias noted that they don't update the qualification every release, maybe every 6 months. Also when there is a new release of a package, a new dependency could cause an issue, but they permit it to pass because of knowledge about the package and the introduced risk is low. Hence there is no clear cut rule but have to use common sense based on the data available

Andy agreed that the risk score should be used as indication only since a brand-new package by Hadley Wickham with only 4 functions may score medium/high risk as it's new, but be classed as low risk because of the knowledge we have about the author.

Manage Requirements / Tests Stream update

Nate Mockler, Keaven Anderson and Tilo Blenk have formed a new stream to look at the requirements and tests. The team met & discussed some testing that Tilo has done, but it was more quick tests at the time of install. Keaven has questions of what the team should be discussing as we need to ensure they are not duplicating the work in the Metrics stream. There is a slack channel where people can add their comments – please provide feedback so the team know what to focus.

Keaven will share the teams notes with the group to get feedback on what the should be working on / developing. The team are unclear what the best value to get out of the team.

Some question we have are:

- What level of testing is needed.
- How do you test that the inputs are valid.
- How much do you cover key functions vs nuisance functions.
- How do you document tests for someone who wants to use the package.
- It doesn't have 80% coverage should we still look at tests?

Website stream updated

Andy has updated the website so that it now has an overview page to introduce validation concept but with main focus on the Validation Page itself. There is still work to do and we need to ensure that someone new to the project can really understand what we are doing straight away. It was felt that more overview text is needed to achieve that. Andy will start the other items on Slack for discussion

R “Validation” Hub Meeting

02 April 2019

On the “Who are we” page, the preview site now has company names and our names and will include Logo’s. Therefore if anyone does not want their name on the main website please let us know immediately. No emails will be shown, just names so that if someone from your company reads the website, they know to talk to you about it. It will be clear its individuals contributing to the project and not companies endorsing.

Everyone was encouraged to participate in the discussion on Slack, since these meetings are short and more discussion can happen through the Slack channels to achieve further progress between meetings.

Actions

Action Item	Assigned team member(s)	Deadline	Status
Communication, dissemination of information about the project to the wider group Possible Funding discussion if we need to write a package to pull metrics Will re-assess for the next Autumn application	Andy Nicholls/ Lyn Taylor/ Joe Rickert	Sept 2019	Ongoing
Review of the website content: Direct review back to Andy, via GitHub or message via slack as you prefer.	Kieran Martin, Min Lee, Juliane Manitz, John Simms, Kieven Andersen, Alex lock-Achileos	NA	Ongoing
Risk assessment workstream – Continue work on objectives	Yilong Zhang , Rebecca Krouse, Doug Kelkhoff, Matthias Trampisch, Eric Nantz	May 2019	Ongoing
Requirements/tests workstream set objectives + milestones	Keaven Anderson , Nate Mockler, Tilo Blenk.	May 2019	Ongoing
Validation white paper workstream – appoint spokesperson and set objectives + milestones	Andy Nicholls, Paulo Bargo	May 2019	Ongoing
Write text for the overview website page and request review by website review team	Min Lee (and then Kieran Martin, Juliane Manitz, John Simms, Kieven Andersen, Alex lock-Achileos to review)	May 2019	Ongoing
Write blog post on gathering metrics and using these in a risk assessment	Doug Kelkhoff	TBC	Open
Investigate how to educate people in the structure of R packages. Yilong suggested including on the website an overview of the structure – so if someone has high level questions we can direct them to the information. Andy added this could be in the form of a FAQ page.	TBC	May 2019	Ongoing