Case Study: Risk Assessment of R Packages at Merck KGaA/EMD Serono

Juliane Manitz, Stefan Pinkert, Martin Gregory and Francois Beckers

Version 1.0 - 14 February, 2022

Introduction

Like many other companies, Merck KGaA/EMD Serono has embarked on their journey to enable the use R for regulatory submissions. Following the framework introduced by the R validation hub (Nicholls et al., 2020), we started to develop an algorithm to qualify a CRAN package as a Merck standard package in our GxP environment.

In a nutshell: Given the R Foundation's effort to ensure the validity of base and recommended R packages, these packages are classified as level 1. If an additional R package passes the installation qualification and successfully executes available tests, the package will be made available to the user and (temporarily) classified as level 3 package. Then, an automated risk assessment of R packages is performed based on the test coverage score (more is better) and the riskmetric score generated from the meta-information (smaller is better). If pre-defined thresholds are fulfilled, the package is qualified as Merck standard package (i.e., promoted to level 2), otherwise an explicit (manual) risk assessment is needed. This 3-tier model provides a useful framework for the users to define a risk-based quality control of outputs when using R.

In this document, we introduce our pathway to a risk-based assessment of R packages at Merck. We provide relevant details on the statistical analysis which led to the definition of thresholds supporting a robust classification of CRAN packages as Merck standard packages. We want to inspire other companies and seek feedback from the community.

Merck Validation Framework

The assessment of R package accuracy is part of the process of validation to ensure quality output of statistical analyses. Validation is "establishing documented evidence which provides a high degree of assurance [accuracy] that a specific process consistently [reproducibility] produces a product meeting its predetermined specifications [traceability] and quality attributes" (see FDA's Glossary of Computer System Software Development Terminology). While focused here on R, the proposed framework can be generalized to other programming languages (e.g. Python, SAS, . . .).

The Merck Validation Framework classifies external CRAN packages into three levels of confidence in the accuracy, reliability, and trustworthiness of their functionalities:

1. Core CRAN Packages which are generally accepted to be accurate based on published documentation by the R Foundation
2. Merck add-on standard packages which have sufficient documented evidence establishing trustworthiness.
3. Other R packages for which the user is expected to ensure proper quality control and respective documentation that the specific package functionality results in the accurate outcome. Respective requirements vary depending on the purpose and complexity of the application.

Risk Assessment Algorithm of R Packages from CRAN

The proposed automated risk assessment of R packages is based on a combination of the test coverage and riskmetric score. A process overview is provided in the Figure 1.

If an R package passes the installation qualification and successfully executes available tests (internal and add-on, if applicable), the package will be made available to the user at level 3. Then, an automated risk assessment of R packages is performed based on the test coverage score (more is better) and the riskmetric score generated from the meta-information (smaller is better). If pre-defined thresholds are fulfilled, the package is qualified as level 2, otherwise an explicit (manual) risk assessment is needed.

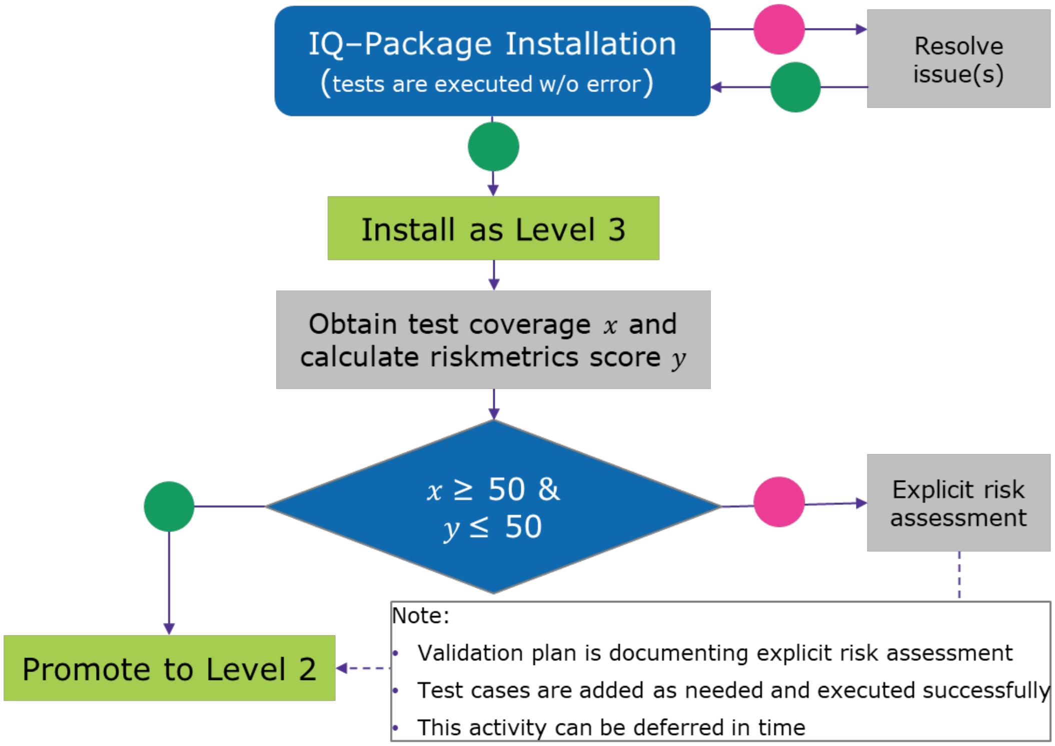
IQ–Package Installation
(tests are executed w/o error)
Resolve issue(s)
Install as Level 3
Obtain test coverage x and calculate riskmetrics score y
$x \geq 50$ & $y \leq 50$
Explicit risk assessment
Promote to Level 2
Note:
Validation plan is documenting explicit risk assessment
Test cases are added as needed and executed successfully
This activity can be deferred in time

Figure 1: Outline of the process for installing CRAN packages in the computing environment.

Riskmetric Score

The riskmetric score has the following components and weights:

- 50% code coverage: unit testing, examples, vignette
- 15% good software development practices: maintainer, public code base, news file
- 15% bug resolution: url, status
- 10% community usage: downloads
- 10% usability metrics: documentation, help, vignette

Although unit test coverage and the riskmetric score are not independent, the overall score has been found to be robust.

Empirical Evaluation

We established a robust threshold for the riskmetric score based on a ROC analysis, which determined optimal classification given the continuous riskmetric score (see Figure 2). As training data, we used a selection of $n = 61$ packages (38 packages were classified as level 2, and 23 packages were classified as level 3).

We find an appropriate threshold for the riskmetrics score at $y = 50$, which is results in a good classification performance (Accuracy = 77% $[64; 87]$). In order to increase specificity, we added test coverage as second dimension with pre-defined threshold of $x = 50$. This results in an improved classication specificity of 88.5%.

Note that as a first version of this automated risk-assessment of R packages for level 2 qualification, we chose a quite conservative approach which deemed acceptable in the general process surrounding the analysis of clinical data.

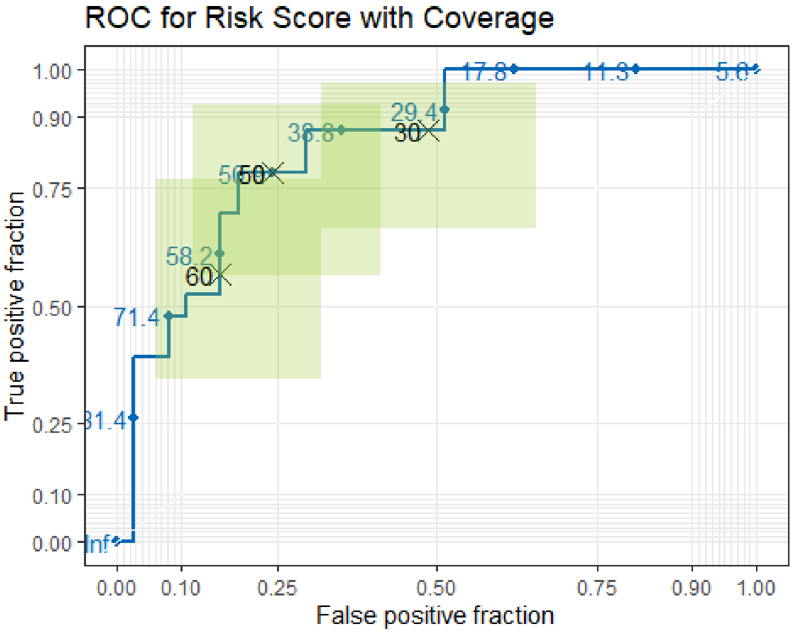

Figure 2: Derivation of Classification Threshold using ROC analysis

Summary and Outlook

We introduced a first version of a risk-based assessment of R packages at Merck KGaA/EMD Serono. The automated risk assessment of CRAN packages classifies R packages based on a two-dimensional risk score, which is composed of test coverage and riskmetric score. The approach results in a final classication specificity of 88.5%, however the accuracy estimates are empirical and associated with some level of uncertainty. Evaluation of a test set of packages and their analysis for potential improvement of the threshold are underway.

We are actively seeking feedback. Please do not hesitate to reach out to juliane.manitz@emdserono.com, and lets discuss during the next meeting of the R validation hub (TBA).