Paper ET12

ValoRpacks - self-service R package validation

Marc Jantke, Entimo AG, Berlin, Germany

ABSTRACT

The use of open-source technologies like R for regulatory submissions is advancing rapidly. While many biopharma companies contribute to the provision and maintenance of high-quality specialized R packages (e.g. Pharmaverse, Tidyverse), their computer system validation (CSV) still remains a cross-industry challenge. The R Validation Hub provides guidelines and tools, but R package validation is a repetitive, resource- and cost-intense company-specific task. GAMP 5 provides clear guidelines on risk-based CSV, which can be adopted for R package validation. Entimo has embarked on a project towards AI-supported R package validation for sponsors and CROs. The project aims to put the ability for package selection, validation and maintenance into the hands of the people who require it for their clinical trial analysis. The project includes a holistic set of guidelines, documents and tools enabling biometrics groups to perform low-effort, yet risk-based and fully regulatory-compliant R package validations from selection to approval.

INTRODUCTION

The use of R in clinical data science has evolved from exploratory use cases into the area of regulatory submissions. Regulatory authorities have confirmed that the use of R for submissions is acceptable, as long as the utilized processes and tools are compliant with applicable regulations and the produced results are reliable, traceable and reproducible. The R Consortium, a non-profit organization, has exercised a series of 5 pilot projects together with the FDA to establish a working model for eCTD submissions with R over the last years and shared the progress and learnings with the industry (1). Several pharma companies have successfully submitted clinical trial data and analysis results produced in R to the FDA (2). There is a common understanding that the use of R in regulatory contexts requires the validation of R packages used for creation of any deliverables, yet the procedure of R package validation is not standardized.

GAMP 5 (Good Automated Manufacturing Practice) defines guidelines how software must be developed, validated and maintained when used in regulatory compliant clinical research and development. These guidelines are known and followed by many commercial providers for software used in clinical trials. R and R packages classify as software according to GAMP category 3 and must fulfill the GAMP requirements for computer system validation of this category. Existing communities are maintaining tools to assist in very specific aspects of validation such as risk assessments (3), traceability matrices (4) and testing (5) or providing white papers describing the validation approach for R packages (6). Yet, a fully functional toolkit and framework to validate R packages is not available.

Pharmaceutical companies and CROs utilizing R for submissions currently have two options to manage and validate the required R packages.

Companies could build a proprietary framework to manage R packages and validations from scratch. This comes with a tremendous ramp up effort and associated cost and puts companies into the responsibility to maintain this framework for eternity (or at least close to, given the retention time for clinical trial submission data and associated tools). Once the framework is operational these are used for the validation of R packages. Typically one version of one package at a time, leading to high operational cost to manage growing package repositories with many packages and versions.

Alternatively companies could purchase R package validation kits from commercial providers. While this is easily accessible, it implies repetitive cost for further packages and package versions. It also creates a dependency to providers with limitations on which packages and versions can be used in regulatory settings depending on the availability. Subsequently R loses one of its main advantages over proprietary software like SAS. The latest statistical functions and procedures are available on CRAN but not ready to be used because the validation of these package versions is solely controlled by the provider.

Entimo is running the ValoRpacks project with the goal to provide the pharmaceutical industry with a self-service framework and toolkit to standardize and automate the GxP compliant validation of R packages for biometrics teams of any size, no matter how far the R adoption has progressed. Entimo is aware of this need from many conversations throughout the industry and conducted a thorough market analysis. The official project was launched in 2024. Throughout an extensive evaluation and assessment phase, entimo has worked on the validation framework and standard templates and is now implementing the individual building blocks of the framework. Entimo is planning to

provide a first version within 2026 to the market and is currently looking for interested companies to cooperate in a phase as early adopters.

PROCESS AND FRAMEWORK

ValoRpacks can be described as an opinionated framework and toolkit for self-service R Package validation. This framework can be used as part of an existing landscape of SOPs and systems, like a validated Statistical Computing Environment (SCE) utilizing the validated R packages, or operated as a slim stand-alone framework for the validation process only.

The framework adheres to the following core principles:

GAMP 5 compliance

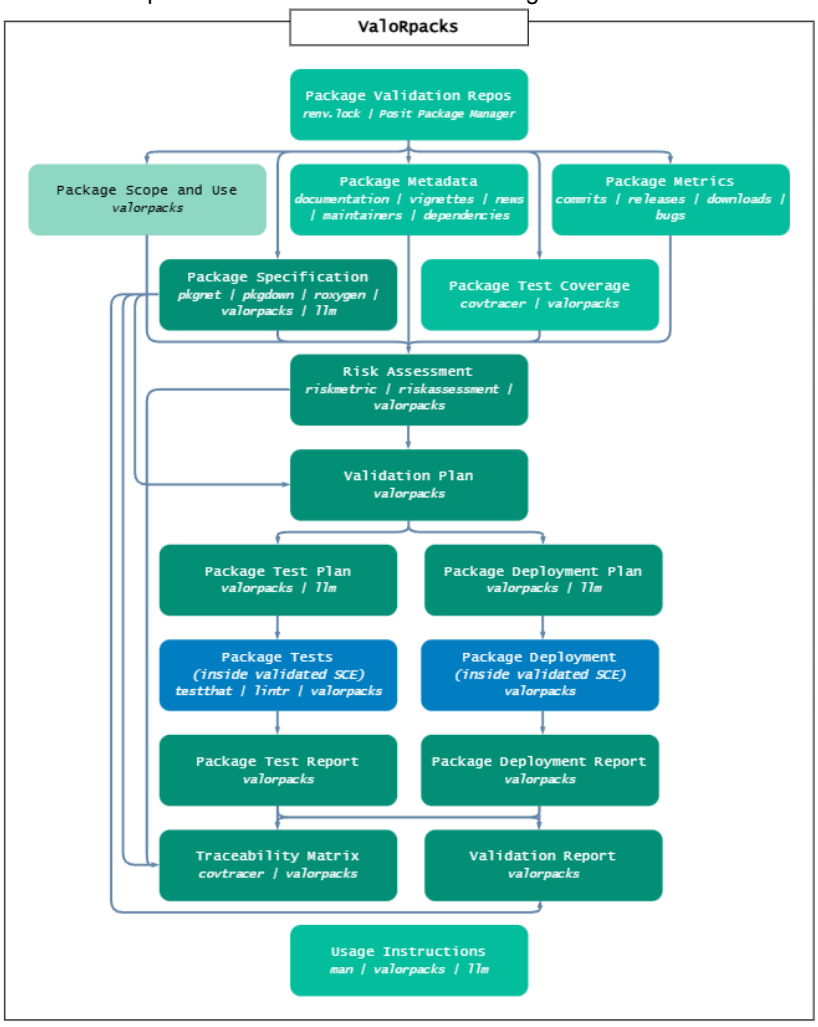
The validation of R packages with ValoRpacks is following the GAMP 5 guidelines for non-configurable software of GAMP category 3 by the book. The validation is driven by a risk assessment and produces comprehensive and traceable validation artifacts.

State of the art tools

The open source community is providing several relevant and established tools supporting various aspects of the R package validation such as risk assessments, traceability and tests. ValoRpacks makes use of these de-facto standards as part of the framework.

Sensible integration of AI

Modern LLMs and Agentic AI is used in ValoRpacks to generate natural language texts for tests, plans and reports as well as to automate the validation process with sensible human oversight.



ValoRpacks utilizes one or multiple separate R package repositories to maintain independent copies of the R packages which subject to validation and retrieved from publicly available repositories, such as CRAN. The user initiates the validation of one or more R packages from this repository and has to specify the scope and use for the selected packages and eventually define restrictions on the page use.

ValoRpacks generates a complete functional specification for the R package to be validated based on available information in the package documentation and source code with the help of available tools such as pkgnet (7), pkgdown (8) and roxygen (9). LLMs assist in producing natural language texts for the specification where needed. Available metadata in the package documentation and from CRAN as well as package metrics from CRAN, github and other sources are used in combination with the defined scope and the generated specification to automatically generate a holistic risk assessment utilizing riskmetric (10) and riskasessment (3).

Specification and Risk Assessment have to be reviewed and approved by humans as a one time effort per package. Once approved, ValoRpacks generates a Validation Plan, Package Test Plan and Package Deployment Plan based on standard templates. Lintr (11), Testthat (5) and available tests will be utilized for the package test coverage. In case the risk assessment demands further tests for specific package functions, these will be generated by ValoRpacks and reviewed by humans. Again LLMs assist in natural language text generation where needed.

Since the selected R Packages have to be validated in conjunction with a validated target Statistical Computing Environment (SCE), the deployment and test execution has to be performed on this target system. The level of automation of deployment and test execution depends on the type of integration with the SCE. Deployment and Test execution generate reports which will be returned to the ValoRpacks framework. Based these reports, the specification and risk assessment ValoRpacks generates a holistic Traceability Matrix and the Validation Report based on further standard templates. ValoRpacks also generates a standardized usage instruction document based on available package documentation in conjunction with the generated specification according to the intended scope and use for the R package.

Agentic AI is used to automate the whole framework process and information management. All validation artifacts are preserved inside ValoRpacks associated to the specific R package version, maintaining the validated state and enabling smooth revalidations for newer versions of validated packages. The framework scales efficiently and is ready to automatically trigger revalidations when new package versions are detected and automate the validation of complete package sets at once.

COMPARISON AND BENEFITS

As outlined in the introduction, companies looking for solutions to the challenge of R package validation currently have two further options besides ValoRpacks to either build and maintain a proprietary validation framework from scratch or to outsource the whole topic of R package validation to an external service provider. While all approaches eventually lead to the same result of having the desired R packages in a validated state available for statistical analysis and reporting in regulated clinical trials, the road towards these results can be quite different

ValoRpacks and any proprietary solution are self-managed and offer full control on which packages in which versions can be validated at which time, only limited by available resources. Relying on external providers on the other hand also means to outsource responsibility and planning, often with the effect that the availability of validation artifacts for specific R package versions cannot be controlled by the company. While the validation of a single package in a single version is a comprehensive task, reality is typically different and much more complex because companies require potentially different frozen package sets and repositories consisting of dozens or even hundreds of specific packages in specific versions which should be available as a validated unit. This is covered in ValoRpacks by design but easily can lead to tremendous additional cost in proprietary solutions or when requested by external providers.

Software validation always requires a qualified target environment, in the case of R packages this would be an R enabled Statistical Computing Environment. Deployment and testing of R packages is an essential part of the validation that must be performed directly on this target environment. Where ValoRpacks offers a direct integration with entimICE FastTrack and an OpenAPI to integrate with other SCEs, this has to be handled individually for proprietary frameworks and is completely impossible when receiving validation packages from external providers.

An often underestimated challenge is the required long term maintenance and support for the R Package Validation framework itself. The framework is used to validate software used in GxP regulated clinical data science activities. Thus the framework is subject to the same level of compliance regarding the software development lifecycle, including maintenance and support. For ValoRpacks and external providers this is in the responsibility and handled by the provider and thus taken care of. Companies building proprietary solutions often do not properly account for infrastructure and personnel requirements to support and maintain the framework over a period of many years, potentially decades.

Aspect	ValoRpacks	Proprietary Framework	Commercial Provider
GxP compliant R Package Validation	X	X	X
Automated standard process with low human effort	X (by the use of automation and AI)	- (manual validation)	- (not relevant)
Latest package versions covered	X (enabled by short validation cycles)	X (if considered in the framework)	- (provider controls R package versions and timelines)
Embedded in Statistical Computing Environment	X (natively with entimICE FastTrack, others as custom extension)	- (not with commercial SCEs)	- (must be handled by the client)
Support for package sets and repositories	X (natively at no additional cost)	X (with additional cost)	X (with additional cost)
Support for multiple package versions	X (natively at no additional cost)	X (with additional cost)	X (with additional cost)
Long term maintenance and support of the validation framework	X (by entimo)	- (typically not considered)	X (handled by provider)

CONCLUSION AND OUTLOOK

The validation of R Packages is getting more important with the increasing adoption of R in regulated clinical research. While the applicable guidelines (GAMP) are clear and useful tools are available, there is no solution available that solves this challenge for companies of all sizes and at all stages of the R adoption journey.

ValoRpacks fills this gap by offering a unique approach to master the validation of R packages. The framework empowers companies to take full control of the validation process and the management of validated R packages and repositories inside their own individual landscape. The high level of automation paired with a sensible level of human oversight enables companies to validate, revalidate and utilize R packages with low cost and low effort.

entimo is planning to release ValoRpacks to the industry in 2026 and looking for cooperations with interested parties and early adopters.

REFERENCES

(1) https://rconsortium.github.io/submissions-wg/pilot_background.html (last accessed 08.01.2026)

(2) https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/EU/Strasbourg/PAP_SA02.pdf | https://posit.co/webinar/novo-nordisk-journey-to-an-r-based-fda-submission/ | https://r-consortium.org/posts/johnson-and-johnsons-success-story-a-hybrid-r-sas-strategy-for-fda-submission/ (last accessed 08.01.2026)

(3) https://pharmar.org/riskassessment/ (last accessed 08.01.2026)

(4) https://github.com/genentech/covtracer (last accessed 08.01.2026)

(5) https://testthat.r-lib.org/ (last accessed 08.01.2026)

(6) https://pharmar.org/white-paper/ (last accessed 08.01.2026)

(7) https://github.com/uptake/pkgnet (last accessed 08.01.2026)

(8) https://pkgdown.r-lib.org/ (last accessed 08.01.2026)

(9) https://roxygen2.r-lib.org/ (last accessed 08.01.2026)

(10) https://pharmar.org/riskmetric/ (last accessed 08.01.2026)

(11) https://lintr.r-lib.org/ (last accessed 08.01.2026)

CONTACT INFORMATION

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

Author Name: Marc Jantke
Company: Entimo AG
Address: Stralauer Platz 33-34, 10243 Berlin, Germany
Email: mja@entimo.de
Website: www.entimo.com