Paper TT13

InnerSource: A stepping stone towards open source in statistical programming

Jeppe Rømer Juul, Novo Nordisk A/S, Copenhagen, Denmark

ABSTRACT

Open source development is a collaborative and flexible way of producing code and documentation, typically with faster iterations, build-in version control and community-driven support. InnerSource is a software development strategy that applies the same practices, but the code is shared only within an organisation.

At Novo Nordisk biostatistics we have transitioned into InnerSource development for almost all programming tasks, including ADaM and TFL programming on individual clinical studies, as well as for tools, apps, R-packages and template programs shared across all projects. Adapting the open source culture has increased knowledge sharing and improved code documentation and support. We believe that InnerSource can serve as a stepping stone towards full open source development in statistical programming, enabling better collaboration and innovation within and beyond the organization.

INTRODUCTION

Traditionally, statistical programming in the pharma industry has not adopted the open source practices, which can present challenges. Programs are often developed in silos, with different teams working on similar projects without sharing code or expertise. This can result in several versions of the same program across projects, leading to inefficiencies and inconsistencies. Moreover, programming tools developed in one department are often not quickly adopted in other departments. In cases where tools and programs are shared across the company, documentation is frequently not kept up to date, making it difficult for other programmers to use the code. Updating shared code can be a long and complicated process, even for a simple and non-controversial improvement. Finally, limited support for tools shared across projects can make it challenging for programmers to work efficiently, resulting in delays and errors in statistical programming.

Open source development is a collaborative approach to programming, where the code is made publicly available for everyone to view, use, and suggest improvements. This approach is characterized by a community-driven approach, which fosters faster iterations, built-in version control, and community-driven support. Moreover, open source development encourages knowledge sharing and innovation, as developers can learn from each other and build on each other's work.

While open source development offers several benefits, it can also present challenges, particularly in the pharma industry. For instance, there may be concerns around intellectual property rights, privacy, and confidentiality of data, which can make it challenging to share code and collaborate openly. Additionally, regulatory requirements and compliance standards can impose additional barriers to implementing open source development in pharma. Furthermore, the pharma industry is often highly competitive, with companies reluctant to share their knowledge and expertise with others.

Despite the challenges associated with open source development in pharma, there is a growing recognition that these obstacles should be overcome to fully realize the benefits of collaboration and innovation. However, until these obstacles can be effectively addressed, InnerSource may be a good approach for companies to still gain the benefits of open source development, while keeping the code and data within the organization. InnerSource can enable knowledge sharing, improved code documentation and support, and increased collaboration within and beyond the organization. By implementing InnerSource, companies can develop the necessary skills and culture to eventually transition to full open source development, while still maintaining control over their code and data.

The rest of this paper provides an overview of how InnerSource is used in statistical programming at Novo Nordisk Biostatistics, and how InnerSource can act as a stepping stone towards full open source development.

INNERSOURCE IN STATISTICAL PROGRAMMING IN NOVO NORDISK

At Novo Nordisk Biostatistics, the principles of InnerSource development are now applied to almost all clinical studies and tools used across projects. This represents a recent shift in the organization, as only three to four years ago, code was neither version controlled or easily readable by anyone outside of the study team.

For study-level programming, each clinical study is allocated its own repository in an AzureDevops project called *Trial Execution*. This repository includes the ADaM programs, macros, programs performing statistical analysis, TFL programs, and certain metadata and utility programs, e.g. to create define-xml. Programmers working on a study can clone the repository to their own local workspace, make changes, and commit and push them to the central repository. This triggers a pipeline that releases the updated programs to the correct place on the UNIX server where the data resides. By giving each clinical study its own repository, it is easy to track changes and maintain version control over the code. Additionally, programmers working on other studies can read the programming done on all other studies, even though they may not have access to the individual subject level data due to confidentiality. This way, knowledge transfer can occur across studies, enhancing collaboration, innovation. Furthermore, programmers from other parts of the organization can suggest code updates in the form of pull requests, if, for instance, a bug has been found in a similar program in their part of the organization.

A number of programming tools, apps, and utilities are shared across projects in Novo Nordisk. For example, the UNIX script "trace" is used for signing programs to document program review, while the Perl script "mkresults" compiles tables and figures displaying the main conclusions on a study into a single document, which can be presented at a result meeting or sent to stakeholders. Furthermore, an ecosystem of R-packages is used across projects when reporting clinical studies in R. These tools, along with many others, are now developed in an InnerSource manner.

Each tool or R-package is allocated its own repository in an AzureDevops project called *Biostat Open Source*, ensuring version control. When a programmer pushes a change to the repository, it triggers a pipeline that may include automated tests that need to pass and/or a manual review from one or more members of the maintainer group for the tool. Furthermore, there may be a requirement to link the update to a work item in the backlog linked to the tool. If all tests passes and all requirements are fulfilled, the pipeline releases the updated version of the tool to the production environment. This approach enables tools to be developed collaboratively and used across projects to promote efficiency and consistency.

Documentation of each tool is stored in markdown files in the same repository as the tool itself. The pipeline renders the documentation to an HTML file placed next to the tool in the production environment, and furthermore it publishes it to a common wiki on the AzureDevops project. This ensures that an overview is maintained of all documented tools, helps promote knowledge sharing and collaboration across projects.

Support for both tools and workflows is done through StackOverflow for Teams. This platform allows programmers to ask and answer questions related to each tool, program, or R-package, thereby promoting knowledge sharing and collaboration across the organization. By using StackOverflow for Teams, programmers can quickly find answers to questions or solutions to problems that other programmers have experienced before them, reducing the time and effort spent on troubleshooting. This community-driven support fits well with the community-driven programming approach of InnerSource, as programmers can collaborate and provide support for each other, even across different projects and teams.

INNERSOURCE AS A STEPPING STONE TO OPEN SOURCE

As outlined above, InnerSource has proven to be a successful approach to developing study-level programs, tools, apps, and R-packages that are used across projects within Novo Nordisk Biostatistics. However, transitioning to true open source for most of these tools and R-packages could offer even more benefits. By inviting programmers from outside Novo Nordisk to use the same code and contribute to its development, the code would become more robust and of higher quality. This approach could also help prevent duplicate work across companies, as the code would be available for anyone to use and build upon.

In addition, regulatory authorities would become more familiar with and build trust in code that is shared between several companies. This could help streamline the regulatory approval process, as the code would be well-documented, tested, and validated by multiple organizations. Furthermore, true open source development could foster greater innovation by enabling developers to build on each other's work, leading to the development of more effective and efficient tools and R-packages.

InnerSource is an effective way to prepare for the transition to full open source development. By using open source practices, such as developing code in a Git workflow, programmers become familiar with the tools and processes involved in open source development. Additionally, InnerSource promotes knowledge sharing and collaboration within the organization, building trust in the code and processes used for development. This trust can help to ease the transition to full open source development, as developers become more comfortable with sharing their code and collaborating with others outside their team. By leveraging the benefits of InnerSource, companies can develop the necessary skills, culture, and trust to successfully transition to full open source development, promoting greater collaboration, innovation, and efficiency in their software development processes.

CONCLUSION

This paper has explored the benefits of applying open source practices in statistical programming in the pharma industry. Traditional programming practices can lead to inefficiencies, inconsistencies, and limited collaboration, but InnerSource development enables companies to adapt open source practices and promote knowledge sharing and collaboration while still maintaining control over their code and data.

At Novo Nordisk Biostatistics, InnerSource has proven to be a successful approach to developing study-level programs, tools, apps, and R-packages that are used across projects. By using InnerSource, the code becomes more robust and of higher quality, facilitating knowledge transfer across studies and enhancing collaboration and innovation.

Looking ahead, the potential benefits of true open source development in the pharma industry are significant, including reduced duplication of effort, streamlined regulatory approval, and greater innovation. By leveraging open source practices, companies can develop the necessary skills, culture, and trust to successfully transition to full open source development, leading to better outcomes for patients and the broader healthcare community.

ACKNOWLEDGMENTS

The framework for InnerSource development of tools and study-level programs is a result of the combined efforts of the statistical programming community in Novo Nordisk. Special acknowledgements go to Christian Haargaard Olsen, Sia Louise Christiansen, Steffen Falgreen Larsen, Johan Ejstrud, and Ari Siggaard Knoph, whose contributions were instrumental in the success of this initiative..

CONTACT INFORMATION

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

Jeppe Rømer Juul
Novo Nordisk A/S
Novo Allé 1
2880 Bagsværd
Denmark
Work Phone: +45 30750940
Email: jxju@novonordisk.com