Paper OS07

# Building Rome in a Few Months: How Open Source Technologies Empower Small Pharma

Chendi Liao, Recursion, Salt Lake City, USA
Kasra Yousefi, Recursion, Salt Lake City, USA

## ABSTRACT

Traditionally, small pharmaceutical or biotech companies tend to rely on FSPs or CROs for statistical programming and analytics support through the use of commercial software. The emergence and maturation of enterprise-caliber open-source software and tools present new and transformative opportunities by helping to enhance data accessibility, streamline analysis workflows, and foster collaborative innovation. However, building these types of infrastructure within small pharma involves navigating budget constraints, limited IT resources, stakeholder buy-in and the need for specialized expertise. We delve into the practical aspects and lessons learned from our experience of building up an open-source Statistical Computing Environment utilizing Posit Ecosystem. By addressing infrastructure hurdles, the integration of open-source toolsets has significantly improved our data analytics capabilities accelerating drug development. A similar approach can be adopted by other small and mid-sized biotech/techbio companies aspiring to become active users of open-source technologies in life sciences.

## INTRODUCTION

In the realm of small pharmaceutical or biotech companies, traditional practice has seen heavy reliance on Full Service Providers (FSPs) or Contract Research Organizations (CROs) for support in statistical programming and analytics through the deployment of proprietary software. However, recent years have seen a game-changing shift with the evolution of enterprise-grade, open-source tools and software. This progression presents a wealth of transformative opportunities, including enhancing data accessibility, streamlining analytical workflows, and cultivating a platform for collaborative innovation. Despite the seemingly clear advantages, the transition toward open-source infrastructure can be fraught with challenges. This paper discusses the comprehensive aspects and practical lessons learned from our firsthand experience in establishing an open-source-based Statistical Computing Environment (SCE) within a small pharma landscape.

## LIMITATIONS OF OUTSOURCED MODELS

The main drawbacks faced by organizations that depend extensively on FSPs or CROs include high costs for trials, reliance on proprietary software, inflexible processes, lack of expertise in advanced analytics, and slow data-to-insight transformation as outlined below

### High Cost

The financial commitment required for a Phase I trial can be considerable. The typical cost of a Phase I trial ranges from $1.5-6.5 million[1], which can frequently pose a significant barrier for many organizations. This initial investment often becomes a crucial factor in the decision-making process.

### Proprietary Software / Solutions

Dependence on proprietary software and solutions is also problematic. These systems often restrict the freedom of operation and sometimes lead to compatibility issues. Moreover, the cost of acquiring, maintaining, and updating these proprietary solutions can add significantly to the overall operational expenses.

### Lack of Flexibility

Oftentimes, the outsourced model lacks flexibility, especially when it comes to constant data monitoring, modifying outputs, or increasing the frequency of analysis such as SDTM, ADAM, and TLF generation, without delaying timelines or incurring additional costs.

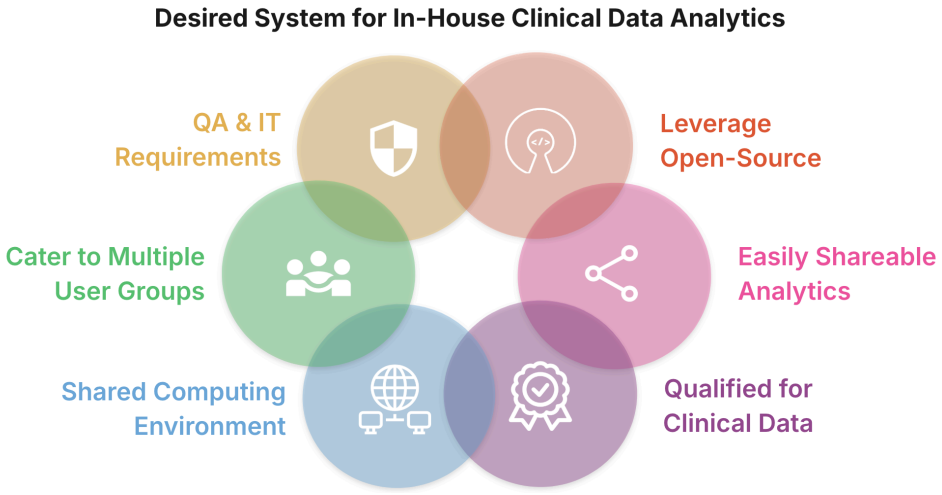
### Lack of Expertise in Advanced Analytics and Visualizations

Analytical and visualization skills are vital in today's data-driven world. However, a lack of these capabilities can impede the ability to make data-driven decisions and derive meaningful insights. There is often a steep learning curve associated with these tools, both for the analytical developers, as well as the stakeholders who view and interpret the results and visualizations. Without adequate expertise, organizations might not utilize them to their full potential.

### Long Time Between Data and Insights

In today's fast-paced world, the speed at which data is turned into actionable insights can be a determinant of a company's success. However, all of the above mentioned issues with outsourcing, in conjunction with process inefficiencies, lack of expertise, or inadequate tools can cause a lag. This delay can slow down decision-making processes and potentially lead to missed opportunities.

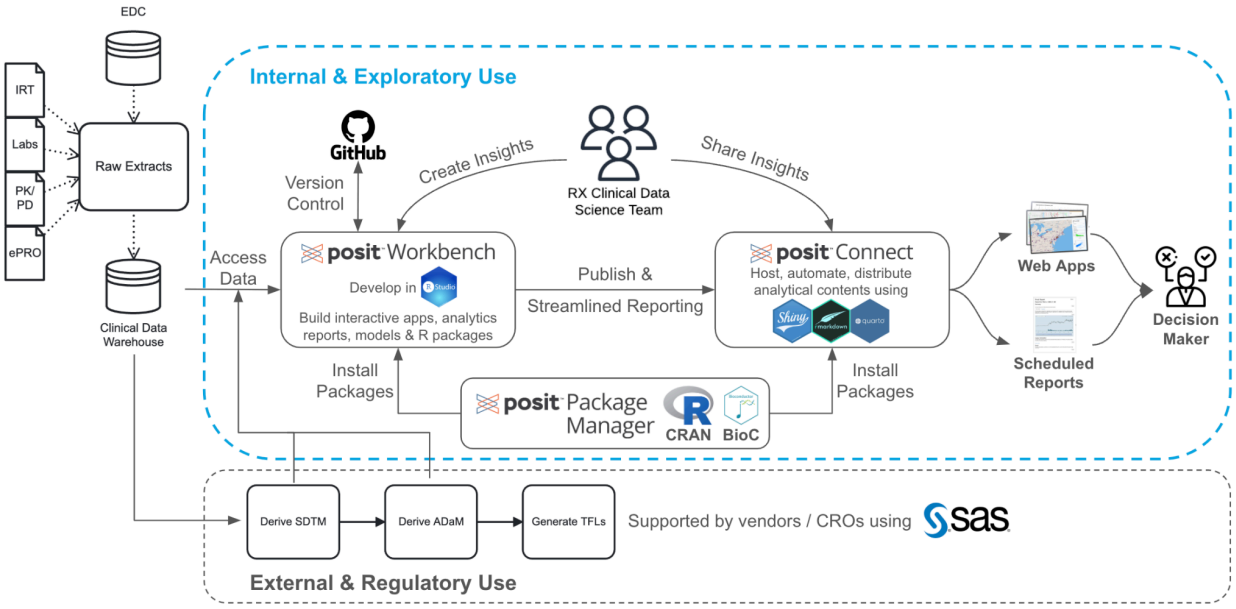
## OPTIMAL INTERNAL INFRASTRUCTURE




An ideal in-house statistical computing environment for clinical data analytics should be built on open-source technologies, enabling flexibility, cost-efficiency, and continuous innovation. By leveraging robust open-source ecosystems such as R and/or Python, organizations can integrate state-of-the-art statistical methods while maintaining transparency and reproducibility. The system must offer intuitive interfaces along with advanced scripting capabilities to support diverse user groups who might interact with clinical data, including not only statistical programmers, but also biostatisticians, bioinformaticians, biomarker scientists, pharmacokinetics/pharmacodynamics (PK/PD) scientists, real-world data (RWD) scientists, etc. It should also facilitate easily shareable analytics, allowing cross-functional teams to generate, review, and distribute insights through interactive dashboards, automated reports, and collaborative notebooks, ensuring efficient knowledge transfer across stakeholders.

Furthermore, to support clinical research and regulatory compliance, the platform must be qualified for handling sensitive clinical data, adhering to validation, security, and audit trail requirements. A shared computing environment should be established to optimize resource allocation, enabling scalable execution of complex analyses while ensuring data integrity and controlled access. Additionally, it should easily align with organization's internal QA and IT requirements, incorporating version control, traceability, and robust authentication protocols to meet enterprise security and governance standards.

Given these criteria and aligning with precedents seen in the pharma industry, we chose Posit Team as the foundation of our clinical data analytics platform. Posit Team provides a robust open-source-based ecosystem that seamlessly integrates R, Python, and other productivity tools like GitHub, enabling flexible and reproducible statistical computing. Its shared computing environment, powered by Posit Workbench and Posit Connect, allows users to develop, deploy, and share analyses through interactive applications, automated reports, and APIs. Additionally, Posit Package Manager ensures controlled management of packages and their sources, facilitating QA and IT compliance requirements. Below we demonstrate a hybrid initial data analytics infrastructure/pipeline leveraging Posit Team for small pharma, for use in the internal and exploratory setting.




As organizations expand, stakeholders within the company can consider scaling up the infrastructure further, creating a validated ecosystem and environment utilizing the same open-source methods for external reporting and regulatory-facing use.

## RECOMMENDED STEPS FOR IMPLEMENTATION OF OPEN SOURCE IN SMALL PHARAM

Starting from scratch with open source technologies can feel daunting, especially in a small pharma setting, where resources and expertise are limited. Committing to that first step is often the hardest, but a phased approach can make the transition manageable. We recommend beginning with a clear assessment of your current data landscape and identifying a specific, well-defined project where open-source solutions could offer immediate value. We also provide some tips and pointers below, to help organizations prioritize objectives and take that first step of the journey towards open source adoption.

### Change Management

In clinical development, especially within Small Pharma, relevant stakeholders such as medical monitors, safety scientists and clinical scientists are typically used to the outsourced model and availability of dashboards under this operating model. It is crucial to have a discussion with organizational leaders about the importance of an internal analytics workflow and how this can be done efficiently through utilization of open source tools. This may often be achieved through demonstration of value using a proof of concept (POC) product (refer to section below).

We would also recommend sharing relevant material from recent conferences and publications to highlight the current industry trends and importance of incorporating open source technologies. Building internal expertise through training and knowledge sharing will also be crucial for long-term success and fostering a culture of open-source adoption within the organization.

### Cross-Functional Collaboration

It is crucial to identify other functional teams outside biometrics that also utilize open source software to gain rapport, such as computation biology, PK/PD, Biomarker and RWD teams. Gaining insights on their usages and preferences can better inform the system design. Further, it is crucial to discuss and share experience with colleagues in the IT, Security, Quality and Engineering team, who will be the main force building the internal infrastructure. In our case, it was helpful to initially launch an R-shiny application to demonstrate the value using a minimum viable product (MVP) process (such as using remote desktop) prior to commissioning Posit as the open source platform. Once the platform was chosen and initial design drafted, the project group also created an agile yet secure workflow and standard operating procedure documents, highlighting the simple operating model for the initial use cases.

### Define POC and Roadmap

One of the key steps is defining a clear and achievable (POC) that demonstrates the system's feasibility, value, and scalability. The POC should focus on a well-scoped use case, such as automating a key statistical analysis or streamlining data visualization, to showcase efficiency gains, which in turn validate technical choices. Alongside this, a scalability roadmap should outline the necessary infrastructure, talent, and governance models required for full

deployment. This includes assessing budget constraints, computing resources, regulatory compliance, and IT support—weighing them against the costs and limitations of a fully outsourced model.

For the POC, we leveraged a milestone readout opportunity for an early phase clinical trial by creating an R-shiny application utilizing `teal` [2] to support real-time exploratory analysis and data deep-dive activities. The tremendous value of the POC brought to the clinical study team spoke for itself. Subsequently, support was garnered all-around and the SCE development roadmap was expeditiously reviewed and approved for further resources and budgets.

#### Leveraging Existing Open Source Community

Open source is not just about the tools and technologies, it's also about the spirit of openly sharing learnings and experiences. Embrace the open source community by leveraging the vast resources, documentation, guidelines and knowledge sharing from pilot projects. For the infrastructure set-up, we recommend utilizing already available, known and tested ecosystem services such as Posit. Partnering with experienced consultants or service providers in the open source space can also be an effective approach to guide the initial implementations.

#### Utilizing Open Source Tools

Similarly, for implementation of data analytics pipeline, we also greatly benefited from open source technologies and the fruits of the `pharmaverse` community [3].

The suite of pharma-specific R packages from pharmaverse provides a powerful, flexible, transparent yet highly streamlined approach to conducting clinical data analysis. Creation of CDISC-compliant ADaM datasets using `admiral` [4] helps automate data derivation and standardization, ensuring consistency and traceability in dataset creation. Subsequently, statistical analyses and visualizations can be conducted using clinical analysis-focused R-packages such as `tern` [5], `gtsummary` [6] or `gtreg` [7], providing an efficient way to summarize patient demographics, treatment safety or efficacy.

Furthermore, in conjunction with tools like `rmarkdown`, we have automated the generation of patient profile reports on a regularly scheduled basis for all ongoing clinical trials, for data monitoring and cleaning, patient progress tracking, and ultimately supporting clinical decision-making. Utilizing R-shiny-based interactive data exploration framework `teal` and selective analysis modules from associated R packages such as `teal.modules.general`, `teal.osprey`, and `teal.goshawk` [8], we deployed fit-for-purpose R-shiny applications that enables dynamic exploration of the trial data, rapid subgroup analysis, and streamlined generation of presentation-ready tabular summaries and visualizations. Such tools empowered study teams to efficiently and confidently explore clinical trial data at readout, accelerating the process of insights generation and key findings communication.

Once the analytical ecosystem is built, it is quite straightforward to extend the use case into other areas of clinical development. For instance, our team developed a fit-for-purpose dashboard for clinical operations intelligence, utilizing data from clinicaltrials.gov to provide descriptive and predictive analytics for patient enrollment.

## CONCLUSION

The use of open-source toolsets in our company has greatly enhanced our data analytics capabilities, accelerating our drug development process by facilitating quicker data handling and management. This same strategy can be beneficial for small and mid-sized biotech/techbio companies. Implementing open-source tools can provide these organizations with cost-effective, flexible technology solutions to improve their processes from research to production. However, it's important to note that the success of this approach requires ensuring team members have the necessary technical skills and resources to effectively use these open-source tools.

## REFERENCES


1. Sertkaya A, Wong HH, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. Clin Trials. 2016 Apr;13(2):117-26. doi: 10.1177/1740774515625964. Epub 2016 Feb 8. PMID: 26908540.
2. Kaledkowski D, Rucki P, Chlebowski A, Verissimo A, Kirar K, Viyash V, Kosinski M, Waddell A, Burkoff N, Hallal M, Nasinski M, Pagacz K, Zhao J, Lewandowski T (2024). teal: Exploratory Web Apps for Analyzing Clinical Trials Data. R package version 0.15.2, https://insightsengineering.github.io/teal/.
3. Pharmaverse. https://pharmaverse.org/
4. Straub B, Bundfuss S, Dickinson J, Farrugia R, Gautier F, Kulkarni D, Mancini E, Mascary S, Miller G, Rothstein J, Shapcott S, Simms E, Sjoberg D, Thoma S, Webster P, Zhu Z (2025). admiral: ADaM in R Asset Library. R package version 1.2.0, https://pharmaverse.github.io/admiral/.
5. Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. The R Journal 2021;13:570–80. https://doi.org/10.32614/RJ-2021-053.

6. Zhu J, Sabanés Bové D, Stoilova J, Garolini D, de la Rua E, Yogasekaram A, Wang H, Collin F, Waddell A, Rucki P, Liao C, Li J (2025). tern: Create Common TLGs Used in Clinical Trials. R package version 0.9.7, https://insightsengineering.github.io/tern/.
7. Pileggi S, Sjoberg D (2024). gtreg: Regulatory Tables for Clinical Research. R package version 0.4.0, https://shannonpileggi.github.io/gtreg/, https://github.com/shannonpileggi/gtreg.
8. Liao C, Unardi D (2024). Revolutionize Clinical Trial Data Exploration with *teal*. PHUSE US Connect 2024. https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/US/Bethesda/PAP_DV04.pdf

## ACKNOWLEDGMENTS

Chris Jung, Recursion, Salt Lake City, USA
John Merrill, Recursion, Salt Lake City, USA
Harry Yang, Recursion, Salt Lake City, USA

## CONTACT INFORMATION

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

Author Name: Chendi Liao
Company: Recursion
Email: chendi.liao@recursion.com

Author Name: Kasra Yousefi
Company: Recursion
Email: kas.yousefi@recursion.com