Paper AD10

The Open SCE: A Call to Action One Year Later

Caroline Phares, Altair, Devon, PA, USA
Pascal Bouquet, HealthTechPartners 42, Basel, Switzerland

ABSTRACT

Last year, a paper entitled, *The Open-Source SCE*[1], was presented at PHUSE EU Connect and included in the open-source webinar series, communicating the community's inception and its mission to develop an open-source toolkit for statistical computing in clinical trials. This paper provides an update, describing the challenges and changes to the community, and aims to inspire stakeholders to join and form a working group focused on developing innovative tools that set new benchmarks in orchestration and auditability. This environment will support collaborative development, session settings management, and the assembly of individual outputs into publication-ready documents, including TLF, ADAM and ARS (Analysis Results Standards) datasets, and data visualization apps. We aspire to advance these tools, enhancing transparency, compliance, and efficiency across clinical studies. We invite software developers, data scientists, regulatory specialists, and others to join us in this pioneering effort to transform pharmaceutical research and development through innovative open-source technology.

INTRODUCTION

Statistical Computing Environments (SCEs) are essential for managing the complex data analysis and reporting workflows in the pharmaceutical industry. As clinical trials and regulatory requirements grow increasingly complex, the need for robust, flexible, and scalable SCEs becomes more critical. Traditional SCEs, often reliant on proprietary software, have faced challenges such as vendor lock-in, limited flexibility, and high costs. The Open SCE initiative was conceived to address these issues by promoting an open-source framework that can foster innovation and enhance collaboration across the industry.

The Open SCE initiative aims to create a flexible, modular, and community-driven SCE that can be easily adapted to meet the diverse needs of different pharmaceutical companies. By leveraging open-source technologies, the initiative seeks to break down the silos that have traditionally characterized the industry, enabling more seamless integration of new tools and technologies. This paper discusses the history and development of the Open SCE initiative, the challenges it has faced, and research conducted through extensive interviews with industry leaders. The interview findings highlight the challenges and opportunities associated with adopting open-source solutions in this highly regulated sector which have inspired the proposed next steps in growing an Open SCE community, inviting broader participation from across the industry to drive innovation and collaboration.

THE OPEN SCE

PURPOSE

The Open SCE is an effort aimed at establishing a lasting, evolving, community-driven, open-source Statistical Computing Environment (SCE) framework tailored to the specific needs of the pharmaceutical industry. Key themes include the transition to cloud-based environments, the integration of AI, the critical need for orchestration and code management tools, and the potential for industry-wide standardization and governance. The mission to standardize and collaborate in the SCE space is not new, and it has inspired many attempts by industry leaders and practitioners, software vendors, and service providers alike to leverage commonalities in order to reduce complexity and effort.

THE PHUSE SCE WHITEPAPER

The *PHUSE Statistical Computing Environment Whitepaper*[2] marked a significant milestone in the pharmaceutical industry's approach to SCEs. It highlighted the critical need for improved orchestration and auditability within SCEs, noting that the industry was moving towards more cloud-based solutions and modern workflows. The whitepaper underscored the importance of these transitions in enhancing data integrity, compliance, and efficiency across clinical trial processes.

The whitepaper also identified key areas where the current SCEs fell short, particularly in their ability to support the rapidly changing technological landscape. It called for the development of more flexible and scalable solutions that could adapt to new challenges, such as the integration of AI and machine learning into data analysis workflows. These findings laid the groundwork for the Open SCE initiative, which seeks to address these challenges by leveraging open-source technologies and fostering a community-driven approach to SCE development.

THE OPEN-SOURCE SCE

Building on the insights from the PHUSE SCE Whitepaper, the Open SCE initiative was launched with the goal of creating a more flexible and vendor-independent SCE framework. Presented at PHUSE EU Connect 2023, the initiative aims to develop an open-source SCE that can be customized to meet the specific needs of different pharmaceutical companies. The Open SCE framework is designed to be modular, allowing companies to integrate the components that best fit their existing workflows and technologies.

One of the key goals of the Open SCE initiative is to foster a more collaborative approach to SCE development. By bringing together stakeholders from across the pharmaceutical industry, the initiative seeks to create a community-driven framework that can evolve in response to the industry's changing needs. This approach is intended to reduce the reliance on proprietary software, which often leads to vendor lock-in and limits the flexibility of SCEs. Instead, the Open SCE initiative promotes the use of open-source technologies that can be freely modified and shared, enabling greater innovation and collaboration.

THE CHALLENGES

Despite its potential benefits, the adoption of open-source solutions in the pharmaceutical industry faces several significant challenges. One of the primary barriers is the lack of broad industry-wide involvement in open-source projects. While there are pockets of innovation, many companies are hesitant to fully commit to open-source solutions due to concerns about the reliability and long-term support of these tools.

Another major challenge is the cultural resistance to change within the industry. Pharmaceutical companies have traditionally relied on proprietary software, which is often perceived as more reliable and secure than open-source alternatives. This perception, combined with the stringent regulatory requirements of the industry, has made many companies wary of adopting open-source tools. Additionally, the complexity of integrating open-source tools with existing proprietary systems can be a significant hurdle, as it often requires substantial investment in time and resources.

The slow pace of industry consortiums and the absence of robust governance frameworks also pose challenges to the adoption of open-source solutions. Without a clear governance structure, there is a risk that open-source projects become fragmented and difficult to maintain, leading to a lack of standardization across the industry. To address these challenges, the Open SCE initiative emphasizes the importance of creating a formal governance framework that can ensure the quality, security, and compliance of open-source tools.

The Open SCE has had additional challenges. SCEs encompass a broad set of functional areas, as was described at a high-level in the *PHUSE SCE Whitepaper*[2], each of which is complex within itself. These functional areas require further explanation and standardization at a low enough level to prompt a focused development effort. One of the ways the Open SCE community addressed this challenge was to take one functional area, such as code management, and begin an effort to standardize on best practices for using git in statistical programming. This proved to be of particular interest to some industry leaders. However, another challenge was that the community did not have enough industry presence, and participation was primarily from software vendors, system integrators and service providers. As a result, although the interest and intent has been present, the motivation has been lacking.

ADDITIONAL RESEARCH

In order to determine if and how the Open SCE community could be helpful to the industry, we conducted a series of interviews with key stakeholders from Sanofi, Novo Nordisk, Boehringer Ingelheim, Johnson & Johnson, and Novartis. The intent of those interviews was to understand their definition of an SCE, the state of their companies' current SCE solutions, their on-going activities to modernize, the gaps and challenges they are facing, and where open-source solutions could help. We encountered varying perspectives on what they define as an SCE. Some companies view the SCE narrowly as the set of capabilities required to generate ADaM datasets and TFL outputs. In contrast, other companies adopt a broader definition that includes capabilities for modeling and simulation, as well as exploration and secondary use of data. Despite these differences, all companies agreed that no vendor currently offers a comprehensive, out-of-the-box SCE solution. What vendors often refer to as an SCE are, in reality, only components of a fully functional SCE. Consequently, there was unanimous agreement that developing a reference model for a complete SCE would be highly beneficial.

These interviews provided also wealth of information on the current challenges, strategies, and innovations related to SCEs. Four primary themes emerged from these discussions: code management, QC management, study orchestration, and the application of AI.

CODE MANAGEMENT

One of the most significant topics discussed across the interviews was the management of code repositories within SCEs, particularly the adoption and implementation of git. Several individuals highlighted the difficulties associated

with using git, particularly for statistical programmers who may not have a software development background. For example, at Sanofi, Andre Couturier and Cindy Marabotti noted that while git is a powerful tool for code management, its complexity poses a steep learning curve for non-developers.

The interviews also revealed varying opinions on how git should be integrated into the SCE workflow. For instance, some individuals suggested that while git is essential for version control, its workflows do not align well with traditional statistical programming practices. Life science companies are considering ways to simplify git for statisticians, possibly by integrating it into their existing one-click version control system. Another challenge discussed was the integration of git with other systems used in the SCE. One company, for instance, is exploring how to align user access management with git as they transition to a more modern solution that combines on-premise servers and cloud storage.

One opportunity for use of the Open SCE community lies in the potential for standardizing git workflows across the pharmaceutical industry. By developing best practices and guidelines for using git in SCEs, the industry can ensure more consistent and efficient code management practices, facilitating collaboration and reducing the risk of errors. This validates the direction in which the Open SCE community was heading.

The need for more user-friendly interfaces for git-based code management was also a recurring theme. All the companies involved in the interviews are exploring options on making git more accessible to statistical programmers. To address this challenge, the Open SCE could explore the development of an open source “simplification layer" that would abstract some of git’s complexities, making it more accessible to users unfamiliar with git. This would aid in the effective adoption of git across different teams within the organization.

QC MANAGEMENT

Quality control (QC) management is a critical aspect of SCEs, ensuring the accuracy and reliability of clinical trial analyses. The interviews provided insights into the current practices and future directions for QC management in the context of SCEs. Because all companies design their QC processes to meet regulatory standards, there is an opportunity to standardize QC best practices across the industry which could increase efficiency.

Traditionally, double programming has been the gold standard for QC in the pharmaceutical industry. However, several individuals who were intrviewed expressed a desire to move away from this practice, describing it as outdated and labor-intensive. Instead, they are exploring the use of AI and automation to enhance QC processes.

As an example, for a mid-size company, QC is managed through a combination of log scanners and validation status checks. Their system includes tools that automatically scan logs for errors, warnings, and critical notes, generating reports that help ensure compliance and data integrity. These reports are interactive, allowing users to drill down into specific issues, helping to streamline the QC process.

The integration of AI into QC processes was a significant focus of the interviews. Companies are exploring how AI can be used to automate QC, reducing the reliance on manual processes and improving efficiency. This topic is further elaborated upon in a separate section focused on AI.

STUDY ORCHESTRATION

The concept of an “orchestration layer” in SCE solutions was a recurring topic across the interviews. The orchestration layer is crucial for managing tasks, automating workflows, and ensuring that various components of the SCE function together efficiently. Most companies reported that they still rely heavily on manual processes for orchestration, often using tools like Excel to manage tasks and study metadata. This approach can lead to inefficiencies and errors, highlighting the need for a more integrated and automated orchestration solution. One individual pointed out that each company has its unique methods for orchestration, reflecting company-specific workflows and processes. This customization is necessary to accommodate the diverse needs of different teams and studies, but it also complicates the development of a standardized orchestration layer across the industry.

The orchestration layer should ideally manage the entire workflow from planning to execution. This includes allocating tasks to the appropriate personnel, particularly statistical programmers, and ensuring that each task is appropriately assigned, tracked, and completed on time. There was broad agreement that the orchestration layer needs to integrate seamlessly with existing tools and systems within the SCE. For example, several individuals discussed the need for an orchestration layer that could integrate the SCE with a study metadata management system for the purpose of automating the study programming plan and task list in the SCE.

Several companies are actively exploring ways to automate and streamline orchestration processes. This could involve the development of new tools or the enhancement of existing ones to reduce reliance on manual methods. Automating task allocation and study planning could significantly improve efficiency and reduce the risk of errors. There was consensus that a standardized orchestration layer across the industry could be beneficial. Such a layer would help streamline processes, enhance collaboration, and ensure consistency in how tasks and workflows are

managed across different SCEs. This led to discussions on the potential for using or contributing to open-source tools for orchestration, specifically how components of the orchestration layer could be developed through community-driven efforts. However, there were also concerns about the effectiveness and adoption of open-source solutions within the pharma industry.

One common topic discussed in the interviews was the use of metadata to drive automation and task management. For example, metadata associated with clinical studies, such as protocol information, data mappings, and analysis plans, can be used to automate the creation of programming plans. Metadata can be leveraged to automate the generation of deliverables, task lists and workflows. For instance, predefined metadata structures can inform the orchestration layer about what tasks need to be completed, who should perform them, and how they should be prioritized. This automation would not only streamline workflows but also ensure that all necessary tasks are accounted for and properly managed. Metadata-driven orchestration can simplify task management by providing clear and consistent guidelines for how tasks should be executed. This approach reduces the likelihood of errors and ensures that all team members are working from the same playbook, enhancing collaboration and efficiency across the SCE.

ROLE OF AI

There was a consensus in the interviews that AI will have a role in modern SCE solutions and could be applied in numerous ways. AI was discussed as a potential tool for automating QC processes within the SCE. We explored the idea of using AI to generate QC programs that could independently validate code written by human programmers. This could eliminate the need for double programming, therefore potentially doubling the efficiency of study teams. Similarly, we discussed the possibility of using AI to generate initial drafts of code based on statistical analysis plans and study metadata. This AI-generated code could then be refined by human programmers, streamlining the coding process and accelerating the time to production for deliverables. In addition to AI-generated code, One company had explored the use of AI for creating synthetic datasets. This synthetic data can be used in various ways, such as pre-testing programs before real patient data is available. The ability to generate synthetic data on demand allows for earlier and more extensive testing of programs, which can lead to more robust and reliable outputs when real data becomes available. The long-term vision discussed in several interviews includes a scenario where AI agents could take over much of the QC and even some programming tasks. In this vision, multiple AI agents could independently verify code, generate synthetic data, and conduct QC, thus reducing the need for human intervention in these processes. This could lead to a future where AI is deeply integrated into every aspect of the SCE, from data handling to final report generation.

In addition to the use of generative AI, we discussed the possibility of integrating AI into the orchestration layer of the SCE. AI could potentially manage task allocation more efficiently by predicting workloads, optimizing resource distribution, and automating routine processes. This would reduce the reliance on manual orchestration methods, such as those currently done through Excel, and create a more dynamic and responsive orchestration layer, capable of adapting to the changing needs of the study as it progresses.

AI could also play a role in improving real-time access to data within the SCE by leveraging metadata. For example, AI could help in automatically linking and analyzing metadata to streamline the process of data retrieval, analysis, and reporting. This integration could significantly speed up data processing and decision-making. There was also mention of the difficulty in navigating data catalogs and the potential for AI to assist in this area. The challenge of retrieving information like patient demographics from large datasets was specifically discussed, with AI being suggested as a tool to streamline these processes.

While AI offers significant potential, the discussion also acknowledged the need for a balanced approach where AI complements human expertise rather than replacing it entirely. For example, AI-generated code or QC processes would still require human oversight to ensure accuracy and relevance. There was also a concern about the potential for AI to introduce bias, particularly if it is used to generate code or validate existing code. Ensuring that AI systems interpret statistical analysis plans correctly and do not replicate human errors was seen to be a critical challenge.

IMPORTANCE OF GXP

Across all the interviews, the topics of GxP compliance and validation were highlighted as critical aspects of an SCE. All companies agreed on the paramount importance of GxP compliance. GxP compliance ensures that data handling, processing, and reporting within the SCE meet the required levels of integrity and reliability. Failure to adhere to these standards can result in significant regulatory and legal consequences. Systems and processes within the SCE must be designed with built-in controls and documentation to ensure compliance with GxP requirements. For example, audit trails, data integrity checks, and controlled access are integral components that help maintain compliance and traceability throughout the data lifecycle.

Validation was discussed as a crucial process for ensuring that all components of the SCE, including software, tools, and workflows, function as intended and meet regulatory requirements. Most individuals emphasized that rigorous validation processes are necessary to demonstrate that their SCE solutions are fit for purpose, especially in environments where patient safety and data integrity are at stake. Validation is not a one-time event but an ongoing

process. As SCEs evolve with new tools, updates, and changes in regulatory requirements, continuous validation becomes necessary to ensure compliance. Companies must establish processes for continuous validation, which include regular testing, documentation updates, and compliance checks. It was a common sentiment that the use of open-source tools presents specific challenges for validation. While open-source software offers flexibility and innovation, it also requires rigorous validation efforts to ensure it meets GxP standards. One company discussed the need for a governance framework that supports the validation of open-source tools, ensuring they are reliable, secure, and compliant with industry standards.

There is growing interest in automating parts of the validation process to improve efficiency and reduce the burden on teams. For instance, automated validation scripts and testing frameworks are being explored to ensure that software updates and configuration changes are fully validated. For example, one individual suggested the Open SCE could help in this regard through a collaborative project with a goal of creating a set of Gherkin scripts for SCE components. This deliverable would both define a common set of SCE requirements that could be shared with software vendors, as well as provide the basis for automated testing of those requirements ensuring full coverage and traceability.

CAPABILITIES MODEL

The outcome of the interviews inspired the development of a draft visual model of capabilities that were commonly identified for inclusion in an SCE. When this draft was presented to some of the individuals who were interviewed, the consensus was that it would be useful in helping the industry to standardize on SCE requirements. It was also thought to be of potential use to software companies, system integrators, and open-source communities in addressing the needs of the industry more effectively.

CONCLUSION

INTERVIEW RESULTS

The interviews revealed both commonalities and differences in how companies approach key components of their SCE solutions. While all companies agree on the need for a more integrated and automated orchestration layer, robust code management based on git, and improved QC processes, they differ in their specific implementations and the overall scope of their SCE solutions. Some view the SCE narrowly, focused on generating datasets and outputs, while others take a broader approach, incorporating modeling, simulation, and secondary data use. Despite these differences, there is a clear consensus on the importance of standardization, the potential of metadata and AI to enhance processes, and the need for a flexible yet unified SCE framework across the industry.

OPEN SCE COMMUNITY RESTRUCTURE

In order to address the common industry needs identified in the interviews, The Open SCE community requires a restructure to allow focus on the key areas identified, as well as broaden participation to a greater variety of companies for true open-source collaboration. Three areas were identified that could immediately be organized into Open SCE community topics, one of which is already a topic of discussion within the existing membership:

1. **Git in Statistical Programming**
   This project will focus first on creating a best practice guide for using git in statistical programming workflows. This would further evolve into development of a git abstraction tool that provides a simplified user interface for statistical programmers to use in managing and promoting their code rather than having to interact directly with git repositories.
2. **SCE Capabilities Model**
   This project will focus first on refining the draft capabilities model to best define the components of a common SCE solution. This would further evolve into focusing on each capability at a more granular level and potentially developing a set of Gherkin scripts to define test-driven requirements.
3. **QC Management**
   This project will focus on harmonization and best practice for QC. This would further evolve into an exploration and proof of concept(s) for using AI to streamline the process.

These 3 projects require leadership to drive the effort, as well as participants who have expertise and opinions to shape the deliverables. We invite anyone who is interested to join the community, and perhaps present on the outcome of these efforts in 2025.

REFERENCES

1. The Open-Source SCE, https://phuse.s3.eu-central-1.amazonaws.com/Archive/2023/Connect/EU/Birmingham/PAP_AD07.pdf
2. PHUSE Statistical Computing Environment Whitepaper, https://lnkd.in/dmjp68ui

ACKNOWLEDGMENTS

Boehringer Ingelheim – Mario Lozina :Mario is a principal data scientist and joined BI in 2007 as a statistical programmer and has been involved with the existing reporting environment and the new SCE environment. He detailed his experience with SCEs and the unique features of the tools he has worked with, emphasizing the need for modern architecture and independence from proprietary software.

J&J Olivier Lecomte: Olivier Lecomte is a senior leader at Johnson & Johnson, Head of Statistical Programming, known for his expertise in statistical computing environments and driving industry-wide initiatives for open-source solutions. He has extensive experience in managing large-scale programming projects and is actively involved in collaborative efforts to enhance statistical computing frameworks within the pharmaceutical industry.

Novartis - Bill Illis: Bill Illis is head of Collaboration & Technology strategy for Novartis Clinical Development & Analytics. He leads the development and implementation of a strategic technology roadmap for the Analytics function, encompassing a start of the art statistical computing environment for the integration, analysis and reporting of clinical trial data in clinical study reports and health authority submissions, good data science practices, and regulatory compliance. He also leads the Digital Data Flow initiative at the Pharma Consortium, TransCelerate, developing industry wide solutions for interoperability based on digital study definitions standards.

Novo Nordisk - Andrew York: Andrew is the business owner for the SCE project at Novo Nordisk. As a clinical data scientist, he serves as a technical advisor to the biostatistics leadership team and chairs the programming experts group. With 38 years of experience, Andrew has extensive involvement in IT projects related to biostatistics.

NovoNordisk - Henrik Lynge: Henrik is the VP Development IT Architecture at Novo Nordisk. Originally a biologist, he spent 18 years in various roles at Novo Nordisk, focusing on business applications for data management and trial management. For the past five years, he has been part of the IT organization, driving agile transformations and modern technology frameworks.

Sanofi - Cindy Marabotty: Based in France, with over 24 years at Sanofi, leads the Analytics and Reporting Technology team. Focuses on automation, visualization, and digitalizing the statistical programming environment.

Sanofi- Andre Couturier: Global Head of Analytics and Reporting Technology at Sanofi, with a strong background in IT and business integration, particularly in clinical development.

CONTACT INFORMATION

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

Caroline Phares
Email: cmphares@gmail.com
LinkedIn: https://www.linkedin.com/in/caroline-m-phares/

Pascal Bouquet
Email: pascal.bouquet@htp42.com
LinkedIn: https://www.linkedin.com/in/pascal-bouquet/