

Objects in Mirror Are Closer Than They Appear – How an Organization Came Together to Accelerate Open Source Adoption

Eileen Ching, GSK, Collegeville, US

Ben Arancibia, GSK, Collegeville, US

ABSTRACT

At first glance, the transition to open source in how we manage and analyze data in drug development seemed distant, obstructed by legacy systems and compliance concerns. Yet, as teams united around a shared vision, the organization discovered that the catalysts for change were already at hand. This presentation explores the journey of innovation and transformation, spotlighting the pivotal role of an open source adoption program in fostering cross-functional collaboration, transparent governance, and technical empowerment. Through clear communications, fit-for-purpose tooling, tailored training, responsive support, and policy updates, many impacted by this transition not only were adopters but also became champions of open source values. Real-world examples illustrate how continuous feedback, risk-based strategies, team adaptability, and sustained momentum helped accelerate the journey from aspiration to achievement. Ultimately, this case study demonstrates that with intent and unified effort, the path to open source adoption is far closer than it appears.

INTRODUCTION

As the pharmaceutical industry evolves toward greater openness, interoperability, and computational innovation, the shift from proprietary analytics environments to open source ecosystems has become a strategic imperative. Across industry forums, including PHUSE, R/Pharma, and CDISC – organizations are collaborating to define modern, standards-aligned, and technology-agnostic workflows for clinical reporting and regulatory submissions. [1][2][3][4]

This case study outlines GSK's strategic initiative, code name STAR, and its accelerated journey toward large-scale R and open source enablement. While grounded in GSK's internal transformation, the insights, pain points, and patterns are highly applicable to any organization undertaking a similar evolution.

When we began the STAR journey, many colleagues described open source adoption as something perpetually “on the horizon” – important, inevitable, but always just out of reach. Yet as we stepped into 2025 with a bold move to position R as our default language for data science and clinical reporting, one thing became clear very quickly: just like any driver knows that “objects in mirror are closer than they appear”, the distance between aspiration and action was far shorter than it appeared.

The capability, the readiness, and for many, the appetite for change already existed across the organization. What we needed was alignment, structure, and a shared belief that we could move faster than we had imagined. This is the story of how an organization – with all its complexity, legacy systems, regulatory considerations, and competing priorities – came together to accelerate open source adoption at scale. And it's a reflection on what we observed firsthand: that once people see the path clearly, momentum builds quickly.

THE VIEW FROM THE MIRROR: WHY OPEN SOURCE, AND WHY NOW

The decision to adopt open source as a key business strategy was not an abstract strategic exercise. It was grounded in the lived realities of our teams:

- Dependence on proprietary languages and analytics environments carry inherent risks such as **licensing costs**
- Our Biostatistics leadership had already committed to “**50% (stretch 70%) open source by end of 2025**” – a bold but necessary direction.
- Across industry, peers were moving quickly, with Roche, Novo Nordisk and others publicizing commitments to **open source statistical computing**. Some had already completed **R based regulatory submissions** and shared their experiences publicly. [5][6][7][8][9]
- Regulatory bodies were signaling increasing acceptance of open source languages and **modern, interoperable formats** like dataset-JSON. [1][3][4][6]

- Across GSK R&D, small but vibrant **R communities** were already experimenting, sharing, and upskilling. Years of grassroots and formal upskilling have created a strong foundation for scale.

In other words, open source capability wasn't ahead of us; it was beside us, waiting to be scaled. We set a clear directive for 2026, all new analyses performed by Biostatistics to be done by R as our default language. A direction like this often feels intimidating when first announced. But as teams leaned in, they quickly discovered that the tools, guidance, and community support they needed were closer than expected.

BRINGING THE ORGANIZATION INTO FOCUS: HOW WE STRUCTURED THE TRANSITION

GOVERNANCE THAT SHORTENED THE DISTANCE

Working closely with the program sponsors and key stakeholders, we established a governance model that emphasized clarity and rapid alignment, which translated to consistent direction and quick decision-making. Our goal was simple: reduce ambiguity, reduce friction, and reduce distance. To deliver change on this scale, we needed more than a program. We needed a coalition.

THE EXEMPTION PROCESS: GUARDRAILS THAT GUIDED, NOT SLOWED

Introduced in 2025, the Open Source Exemption Process formalized when it would be acceptable to not use R/open source as the default language. Some examples include ongoing studies with existing code, high complexity in the transition, and missing open source capabilities. It was accompanied by a defined review and approval process, as well as documentation standards. This risk based approach balanced business continuity with policy enforcement.

Far from being a bureaucratic hurdle, the Open Source Exemption Process gave teams:

- clarity,
- fairness,
- and a structured way to manage risk.

Many teams told us the process actually helped them accelerate, because it clarified when R or Python adoption was expected, not optional.

MEETING PEOPLE WHERE THEY WERE

Although we saw a lot of enthusiasm and excitement, there was also understandable anxiety towards our ambitious goals. Many people have used the same proprietary language and tools for a long time, and some just learned recently. All of them needed personable, responsive, and quality support to be successful on this journey. Enablement was a recurring theme for STAR, mainly in the form of:

- **Quality R environments** – a set of centrally provisioned, controlled R compute environments that differ by how tightly packages are locked down, supporting both GxP and non-GxP work. Frozen environments are fully locked down, contain only assessed and approved packages, ensuring reproducibility and regulatory-grade trustworthiness. With increasing usage of R by broader functional areas across GSK R&D as well as our experiences with maintaining R for GxP environments, new business cases emerged which prompted the idea of a “semi-frozen” environment. Semi-frozen environments add flexibility by allowing installation of “Recommended” packages while still basing everything on the frozen baseline. Finally, “open” environments allow full package installation for non-GxP exploratory work.
- **Tool and Process focused hypercare** – a group of technical experts providing dedicated hypercare support in the form of hands-on, rapid support to unblock users from adopting R, offering real-time mentoring, problem-solving, and practical workflow guidance. In addition to developing R packages to deliver open source capabilities, the R experts on the STARburst Team share insights broadly through daily micro-blogs, examples, and training resources, helping teams adopt modern open-source analytics more efficiently. The clinical programming subject matter experts on the Tiger Team rapidly prototype, document, and deliver workflow solutions, including guidance, training materials, and R tooling, using an agile, code-first, user-centric approach to unblock study teams to accelerate R adoption by building robust, end-to-end R workflows for clinical trial analysis.
- **Grassroots R communities and the STAR Community forum** – various communities and forums that facilitated open, cross-functional Q&A and updates. The STAR Community Forum is a Teams hub where users post R questions, share guidance, and find program updates. It's the first stop to crowdsource answers and tag the STAR team on issues. There were also other sharing sessions and forums organized by individual departments and groups that enabled functionally aligned (e.g., data management, statistics, programming) peer learning and group support.
- **Targeted training and DataCamp rollout** – a combination of training mechanisms to bridge the skills gap. To help with necessary upskilling, we ran tailored sessions for different functional areas with curated learning content aligned to their workflows. In addition, we brought in DataCamp, an interactive, hands-on

learning platform that helps build data and programming skills through browser-based courses, real-time feedback, and curated learning paths. It supports R, Python, analytics, data science, and AI/ML, allowing learners to practice code directly, follow tailored tracks, and develop skills at their own pace.

- **Altair SAS Language Compiler (SLC)** – a multi-language analytics engine that can compile and run SAS, R, Python, and SQL code. It provides a modern lighter weight alternative to the SAS language environment, enabling execution, debugging, and workflow management via tools like the Altair Analytics Workbench or VS Code-based interfaces. It supports data formats commonly used at GSK, provides a mechanism for running SAS programs in combination with open source languages, and offers the flexibility required for the data science ways of working.

Each enabler was designed with one belief: If we make the path clearer, people will move quickly along it.

TECHNICAL TURNING POINTS – WHEN THE HORIZON SHIFTED

Throughout 2025, we saw several pivotal moments where teams collectively realized, “We can do this.” Technical enablement played a key role in ensuring the teams had the fit for purpose and fit for use environment and tools.

R ENVIRONMENTS THAT BUILT TRUST

We had already implemented an internal R4GxP process that laid out a governed path to validated packages so teams could perform regulatory submission work in GxP-suitable, controlled R environments (delivering accuracy, reproducibility, and traceability per the TransCelerate MSA framework) – giving them a way to use open-source packages with confidence. [10] This mirrored industry emphasis on risk-based validation and controlled execution seen in public pilots and sponsor case studies. [3][6]

With STAR launching and scaling, we ramped up the R4GxP pipeline and institutional support resulting in predictable release cadence and rapid increase in the number of packages added to the controlled (or frozen) R environments. Moreover, the introduction of the semi-frozen environments enabled different stages of work, removing all-or-nothing tradeoffs while preserving regulatory rigor where needed. This was a major turning point, combined with hypercare and training, that normalized R as a first-class production option. For our study teams, these R environments were critical for making the shift from “experimental R” to “production R” with greater confidence.

PARQUET: A MODERN FORMAT THAT REFINED EXPECTATIONS

After careful evaluation of several options, Parquet quickly emerged as a future proof standard file format, and it became a tangible symbol of what open source makes possible. It is fully open source and supported broadly across ecosystems such as R, Python, Spark, Hadoop, and ADLC. It is also supported by SAS, including handling the format directly or via conversions between Parquet and Transport (.xpt) files. This makes it ideal for multi-lingual environments and GSK’s open source first strategy.

However, Parquet was not plug-and-play for GSK. Several new components had to be engineered to make it usable across R, Python, and SAS. These included bi-directional tools allowing teams to convert between dataset/file formats, special metadata handling and edge-case logic, application enhancements to ingest Parquet format, and building new open source first tools. The adoption of Parquet showed teams what efficiency, speed, and interoperability could look like. Many described this as the “aha” moment where the value of open source crystallized.

ALTAIR SLC IN PRODUCTION

SAS Language Compiler from Altair offered an alternative to maintaining and executing SAS programs. It helped teams bridge legacy and modern workflows. It didn’t slow migration; it accelerated it by making this “hybrid” approach both possible and safe.

Before production implementation, a multi-phase proof of concept (POC) to assess the viability of this alternative. The POC tested SLC on multiple study types and evaluated it against clear business and technical success criteria, including code compatibility, data access, integration with our Statistical Computing Environment, and performance on large datasets. Following positive POC results, Altair SLC was deployed to production, fully integrated in GSK’s multi-lingual SCE, as a controlled solution to run programs with SAS code while teams continue to adopt R for their new analyses. Since SLC cannot execute all the SAS statistical procedures with all available parameters, we also created a roadmap for closing the statistical procedure gaps, in collaboration with Altair.

PROGRESS TOWARD OPEN SOURCE SUBMISSIONS

When STAR was launched, there was already another program, Nexus, to deliver an interconnected programming ecosystem for clinical reporting that unifies data standards and automates the creation and validation of data objects,

from collection through analysis to regulatory submission, via linked metadata and execution engines. The Nexus tools would be entirely built in open source tooling, therefore the partnership between Nexus and STAR was crucial to the success of both programs. STAR drives capability, adoption, and readiness while Nexus delivers the reporting capabilities that enable R-based workflows end-to-end, including regulatory-compliant and accepted submission data packages.

THE HUMAN ENGINE BEHIND THE ACCELERATION

If there's one theme that I would underscore for other organizations considering similar shifts, it's this: transformation is ultimately a people effort. Collaboration is a common theme throughout the STAR program.

What made STAR work was not only the technical components, but the way teams stepped forward:

- Clinical Reporting Teams setting a goal for >50% code in R (with 70% stretch) by end of 2025, and adopting R and R-based workflows despite tight study timelines and pressure
- Clinical Programming SMEs on the STAR Tiger Team building and sharing practical guidance and tools to unlock end-to-end workflow
- Data Management and Biostatistics partnering on the transition from data acquisition to submissions
- Business and Tech teams collaborating on requirements, deliverables and support
- Our Tech teams responding quickly to environment issues or package gaps
- R communities supporting one another far beyond what any formal training could provide

These were not isolated efforts. They were a demonstration of what happens when people understand the “why,” feel supported in the “how,” and see tangible progress.

User surveys, InfoHub analytics, community forum participation, and pipeline guided improvements shaped our priorities. We made it an important goal throughout 2025 to communicate clearly, frequently, and with gratitude. Our monthly newsletters, STAR Info Sessions, community spotlights, and direct interactions with users and stakeholders all helped to keep momentum high.

WHAT WE ACHIEVED – AND HOW CLOSE WE REALLY WERE

By year end, we saw measurable progress, evidence that our goal wasn't so distant after all:

- 70%+ of clinical reporting programs were already in R.
- 200+ STARburst support requests were resolved, many under tight timelines impacting critical work.
- A consistent, scalable R environment strategy was in place.
- Altair SLC went live in production in our multilingual environment, enabling hybrid and transition workflows.
- Parquet file format gaining real traction across functions.

Looking back, these achievements didn't happen because the path was easy. They happened because, once aligned, teams realized they were already much of the way there. Objects in the mirror were indeed closer than they appeared.

Most importantly, teams across R&D were no longer just “moving to R.” They were owning it – contributing to package development, sharing workflows, and building communities that will outlast this program.

LESSONS FOR THE COMMUNITY

Big change looks far away until you start moving. The greatest barrier is often perception, not capability. We would like to offer the following insight:

1. **You need both ambition and pragmatism** – The timeline must push the organization, but it also must account for regulatory realities, vendor dependencies, and team readiness.
2. **Bridging tools are not shortcuts, they are accelerators** – Tools like Altair SLC do not delay us, they help us move faster with confidence.
3. **Community matters more than documentation** – Community is the backbone of open source adoption. Our communities of practice became one of our most effective accelerators. Formal training matters, but daily peer support changes behavior.

4. **Transparency builds trust** – Increased trust translates to reduced resistance. Clear communication removes fear of the unknown and keeps teams informed. Sharing progress, risks, and roadblocks helped everyone stay aligned through uncertainty.
5. **Momentum is a leadership responsibility** – Every milestone celebrated becomes fuel for the next one, and it needs explicit reinforcement. Every challenge requires clear framing and follow-through.

CONCLUSION

STAR began as a mandate. It evolved into a movement. The path toward open source adoption felt both urgent and uncertain. We knew the “why” was clear: risks associated with dependence on proprietary languages, an industry shifting rapidly toward R, growing regulatory interest, and our organization’s commitment to modernizing how we analyze and deliver clinical data. But even with this clarity of purpose, the size of the effort ahead was undeniable. What we saw was a community learning together, solving problems together, supporting each other, and pushing toward a future where open source tools make our work more transparent, reproducible, collaborative, and innovative.

Our experience leading the STAR program thus far reinforced a truth we have seen throughout our career:

Transformation isn’t about pushing people across distances. It’s about showing them that they were closer than they thought. Now, looking back on the first year, the most important lesson is this: the catalysts for change were already here – our people, our shared ambition, and a willingness to work differently.

What once felt far away – open source submission capability, modern workflows, streamlined tooling – became achievable when the organization aligned around a shared purpose, supported by the right structures and the right people. To others in the industry navigating similar journeys, we offer this:

Your path to open source may be closer than it appears. Your teams may already have the curiosity, capability, and courage to accelerate – they may simply need a clearer line of sight. And once they see it, they will move faster than you expect.

REFERENCES

1. PHUSE Blog: Open Source Technologies – A Year in Review. https://phuse.global/Communications/PHUSE_Blog/open_source-technologies-a-year-in-review-with-phuse
2. Posit Blog: PHUSE US Connect 2024 Keynote – Collaborating Across Pharma. <https://posit.co/blog/open-source-highlights-from-the-phuse-us-connect-2024-keynote/>
3. R Consortium: R Submissions Working Group: Pilot 5 Launch and more! <https://r-consortium.org/posts/submissions-wg-pilot5-pilot6-and-more/>
4. Pharmaverse main site. <https://pharmaverse.org/>
5. Pfizer leaders on open source adoption (PHUSE interview via R Consortium). R-bloggers, 2023. <https://www.r-bloggers.com/2023/11/pfizer-leaders-discuss-the-adoption-of-open-source-predominantly-r-in-clinical-trial-reporting/>
6. Roche: The R Journey to Submission. PHUSE US Connect 2025 (SA01). https://www.lexjansen.com/phuse-us/2025/sa/PAP_SA01.pdf
7. Appsilon Blog: Johnson & Johnson’s 5-Year Shift to R-Based Regulatory Submissions. <https://www.appsilon.com/post/jj-open-source-journey>
8. Appsilon/R-bloggers: Learnings from Novo Nordisk’s First R-based Submission. <https://www.r-bloggers.com/2024/01/learnings-from-the-first-r-based-submission-to-fda-by-novo-nordisk/>
9. YouTube: Novo Nordisk’s Journey to an R-based FDA Submission (Posit). <https://www.youtube.com/watch?v=t33dS17QHUA>
10. TransCelerate Modernization of Statistical Analytics Solutions. <http://www.transceleratebiopharmainc.com/assets/modernization-of-statistical-analytics-solutions/>

CONTACT INFORMATION

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

Eileen Ching

Ben Arancibia

GSK

GSK

Email: eileen.l.ching@gsk.com

Email: benjamin.c.arancibia@gsk.com

Brand and product names are trademarks of their respective companies.