

The Year of Yes! How Saying “Yes” Ignited GSK’s R Adoption

Becca Krouse, GSK, Chapel Hill, US

Ben Arancibia, GSK, Collegeville, US

ABSTRACT (HEADER 1)

Let's be frank, the transition to open-source is hard. To support this transition and GSK's commitment to 50% use of R by the end of 2025, the STAR Program was created. As part of STAR, the STARburst team was launched. The STARburst team was created under the central question of “What would happen if a team of R experts opened a help desk and said “yes” to every inquiry?” This team set out to address immediate challenges in large-scale R adoption through supporting as many Statisticians, Programmers, and Data Scientists as possible. Using a hands-on, collaborative model, the STARburst team directly engaged with users to overcome roadblocks and enhance R integration into daily tasks. But it didn't stop there – to amplify impact, STARburst shared insights on a daily blog, accessible to the entire Biostatistics community. Attendees will learn how this fast-paced, user-focused strategy is shaping future efforts, extending beyond 2025.

INTRODUCTION

Over the past few years, the pharmaceutical industry has made the shift from proprietary analytics environments to open-source ecosystems. Across forums such as 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]

Despite this progress, each organization faces unique challenges when adopting open-source. How can a company successfully move from prototypes, minimal viable products (MVPs), and experiments to full-scaled adoption? This case study describes GSK's support team, STARburst, a focused group within GSK's strategic program (code name STAR) that accelerated the organization's journey toward large-scale R and open-source enablement by saying “yes” to every question and need.

When an organization makes a big strategic choice like this, how do you help users overcome roadblocks, fear, and general resistance to the change? STAR needed a team like STARburst to build trust, resilience, and open-source capability within our teams. This is the story of how a small team of experts used a collaborative, hands-on model to help teams integrate R into their daily work and foster an open-source community.

Can We Do At Least 50% of Our programming in R?

When making strategic decisions, leadership might consult the experts for recommendations. For GSK, however, “Can we do at least 50% of our programming in R by the end of 2025” was not a question, but a leadership directive. The decision to adopt open-source was based on lived realities:

- Dependence on proprietary languages and analytics environments carry inherent risks such as **licensing costs**
- 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.

These insights led our leadership to believe that open-source adoption was not ahead of us, but ready to scale. GSK set a clear directive for 2025 that all new analyses performed by Biostatistics to be done by R as the default language, which was formalized through a strategic program named STAR. A direction like this often feels intimidating when first announced. It was decided that as study teams leaned in, a dedicated group would need to support them by rapidly resolving issues during the transition. The support group was launched as STARburst, built around the central idea of saying “yes” to every inquiry.

Can we use R end-to-end?

Engaging With Requestors

The STARburst team was launched as a 3-month experiment to impact as many users as possible by welcoming all R-related questions to a dedicated mailbox. The goal was simple: to aid in problem-solving while deepening our understanding of user needs and gaps. STARburst was comprised of a group of R experts with a wide range of domain expertise, from clinical programming and statistics to engineering and tool development. STARburst aimed to work collaboratively with users, solving problems *with* them instead of *for* them. This provided a unique opportunity to simultaneously unblock and upskill users on available tools, resources, and techniques for solving problem.

Early in the first three months, the inbox received very simple questions such as:

I am facing an alignment issue in my output. Please find the screenshot below. The values should be in left and the "Number of AE" should be in a single line.

Oftentimes, these questions included a screenshot and no code. STARburst began using these inquiries as opportunities to teach users effective problem-solving skills—encouraging them to share code via GitHub, create reproducible examples (reprex), and consult relevant documentation. Through this shift, it became clear that STARburst was not only answering questions, but also building capability and helping change the culture of GSK R&D.

We also received very broad questions during the first three months such as:

I am reaching out to seek your expertise and guidance on the process of creating Tables, Figures, and Listings (TFLs) in R for my first study. I would greatly appreciate any assistance you can provide to help me understand the workflow and best practices.

These types of fundamental questions revealed friction in early stages of learning, highlighting a need for tangible examples to start from.

At times, users would seek help from GenAI tools, which proved successful for common tasks such as data manipulation using popular packages (e.g. Tidyverse packages). Occasionally, we received questions to our inbox about why a particular code snippet resulted in a ‘function not found’ warning. We quickly learned that, by suggesting invalid code, GenAI can hinder users adopting new or niche R packages. These interactions further highlighted users’ approaches to problem solving and allowed us to understand where GenAI tools are most effective.

Both specific and broad questions provided valuable insights for potential gaps in tooling, workflows, and resources. In all cases, STARburst aimed to provide quick, in-the-moment support to the user while logging the gap for a future robust, long-term solution. During conversations with users, STARburst made a concerted effort to get to the root of the question to ensure users were not struggling unnecessarily. A shift to open-source requires rethinking historical processes and tooling. By asking “why” behind these questions—for example, why a user wanted to replicate a specific, familiar output aesthetic—STARburst sought to challenge the notion of “because we’ve always done it that way.”

Fostering an Open-source Culture

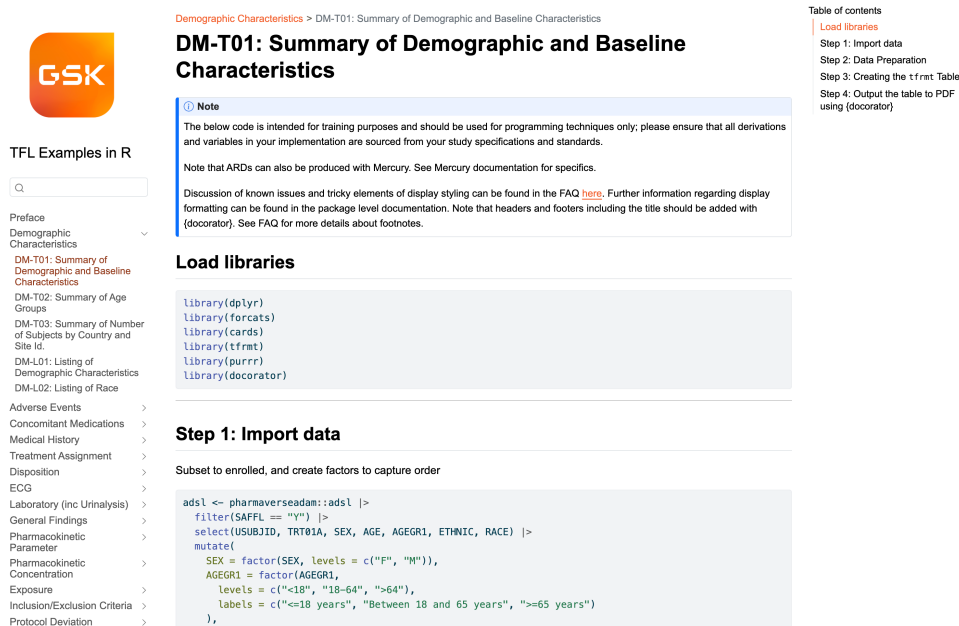
Aside from answering requests directly, STARburst aimed to foster an open-source culture by creating a dialog around R and increasing the accessibility of information. During the first three months of STARburst, our team shared insights and learnings through a daily blog. The blog served to raise awareness of challenges and solutions. In addition to detailing the outcomes of specific requests, the team published general posts addressing common user questions. This approach aimed to share lessons from every query, empowering GSK to build its capabilities and institutional knowledge.

User requests sometimes revealed bugs in open-source R packages or gaps in resources from cross-industry working groups. For instance, many study teams sought assistance with issues during table creation. STARburst helped diagnose the root cause, and if it was a bug in the underlying package, the team fed this insight back to the open-source community. Team members filed issues, discussed problems, and contributed significant pull requests to the popular {gt} package, leading to a release that incorporated many improvements from STARburst. Recent releases of {gt} included many improvements from STARburst team members. [10]

STARburst was not limited to statistical programming needs. The team also contributed to open-source statistical groups like CAMIS - A PHUSE DVOST Working Group. [11] By contributing to CAMIS's website on statistical procedures across programming languages, the STARburst team fed back into a long-standing resource. This ensured that the wider community could benefit in the information, much like GSK itself benefited from the contributions of others.

Applying our learnings in real time

STARburst's interactions with users helped identify key friction points. Early on, we learned that users would benefit from a clear jumping off point. There was an appetite for clear examples for tabling, so in response, we created an internal TFL library of common displays and example code using open-source packages.



Demographic Characteristics > DM-T01: Summary of Demographic and Baseline Characteristics

DM-T01: Summary of Demographic and Baseline Characteristics

Note

The below code is intended for training purposes and should be used for programming techniques only; please ensure that all derivations and variables in your implementation are sourced from your study specifications and standards.

Note that ARDs can also be produced with Mercury. See Mercury documentation for specifics.

Discussion of known issues and tricky elements of display styling can be found in the FAQ [here](#). Further information regarding display formatting can be found in the package level documentation. Note that headers and footers including the title should be added with {docorator}. See FAQ for more details about footnotes.

Load libraries

```
library(dplyr)
library(forcats)
library(cards)
library(tfrmt)
library(purrr)
library(docorator)
```

Step 1: Import data

Subset to enrolled, and create factors to capture order

```
adsl <- pharverseadam::adsl |>
  filter(SAFFL == "Y") |>
  select(USUBJID, TRT01A, SEX, AGE, AGEGR1, ETHNIC, RACE) |>
  mutate(
    SEX = factor(SEX, levels = c("F", "M")),
    AGEGR1 = factor(AGEGR1,
      levels = c("<18", "18-64", ">64"),
      labels = c("<18 years", "Between 18 and 65 years", ">65 years")
    ),
  ),
```

Did It Work – The Outcomes of Saying “Yes”

The short answer is yes, STARburst worked. One of our primary goals in the STARburst experiment was to maintain a daily micro-blog for sharing learnings, tips, and tricks along the way. During the three months, the STARburst team generated 64 total microblogs, with 16,000 views and 821 total GSK viewers of the blog. We received questions from various departments across GSK Biostatistics with 22 from Clinical Statistics, 6 from Non-Clinical Statistics, and 27 from Clinical Programming. They ranged in complexity from table creation to statistical methods in R and parallelization in our SCE (Statistical Computing Environment). STARburst built data products, including our own internal TFL template library based on user needs.

The positive feedback and strong user engagement ultimately led to the decision to transition the STARburst "experiment" into a permanent fixture for the duration of the STAR program.

Can We Scale Our Use of End-to-End R to our R&D Organization As a Whole?

We learned a lot in our first three months of the STARburst program. To truly scale the use of end-to-end R across the entire R&D organization, we recognized that the initial phase, often characterized by examples and quick solutions, had to evolve into a more structured and comprehensive approach. Users were actively seeking more than just isolated examples; they needed clear, holistic workflows that demonstrated how to seamlessly integrate and connect all the disparate pieces of their work. This demand was further complicated by the need to address a multitude of nuances and specific use cases across different teams, therapeutic areas, trial phases, and various regulatory agency requirements. While initiatives like STARburst laid the foundational groundwork, it became evident that additional pillars of support were crucial to move beyond the initial 3-month period and effectively propagate R usage on a larger scale.

This expansion prompted a dual focus: strengthening community and process support. Community building became paramount, leading to regular Knowledge Sharing Sessions (KSS), active leveraging of Subject Matter Experts (SMEs), and increased collaboration through dedicated channels on MSTEams like the STAR Community Forum. In parallel, we put robust process guidance in place by forming a dedicated "tiger team." This team has been instrumental in developing and disseminating detailed workflows for complex scenarios, such as navigating risk-based Quality Control with evolving tooling and managing critical data transformations, like moving from sas7bdat to language-agnostic Parquet for internal file storage. These efforts have provided the foundational support and clear pathways that users need to confidently adopt and integrate R into their daily operations.

Finally, the path to scaling has also involved extending the capabilities of STARburst through interdisciplinary problem-solving. As the use of R has matured and reaches less established processes, problem-solving increasingly requires a multi-squad approach, combining the expertise of STARburst, the tiger team, and the GSK R community to

develop long-term solutions instead of quick workarounds. This has encompassed addressing a variety of statistical needs, refining processes for code and environment submissions, and upholding evolving best practices in coding and package development. We have also continued to evolve our tooling to address identified gaps, particularly in generating Tables, Figures, and Listings (TFLs) and enhancing statistical analysis capabilities, thereby ensuring R remains a powerful and adaptable asset across the R&D organization.

Can we continue to grow in a self-sustaining manner?

For the widespread adoption of R to become truly self-sustaining, a proactive approach to addressing capability gaps and embracing the dynamic nature of open-source is essential. Efforts like STARburst provide hands-on support and continue to reveal fundamental needs, such as guiding users on how to debug their code effectively and manage their R environments, ultimately fostering deeper technical proficiency. Sustained growth hinges on embedding community-driven support as a core principle, building on and expanding initiatives like the daily blog and collaborative problem-solving that are helping to build institutional knowledge within GSK. Recognizing that open-source development inherently involves constant change, our strategy must account for continuous learning and adaptation. This will ensure that the organization not only keeps pace with advancements but also actively contributes to the open-source ecosystem, preparing for a future where innovation is shared and collaboratively maintained.

CONCLUSION

The transition to widespread use of open-source, while challenging, has been demonstrably accelerated at GSK through the strategic implementation of the STARburst team. By adopting a "yes to every inquiry" philosophy, STARburst effectively addressed immediate user needs, built critical R capability, and fostered a vibrant open-source culture through direct support and knowledge sharing. This case study illustrates that moving from experimental prototypes to scaled adoption requires a dedicated, hands-on approach that cultivates trust and resilience. Looking forward, sustaining this momentum hinges on evolving beyond initial problem-solving to establish clear workflows, expand community and process support, and continuously adapt to the dynamic open-source landscape. Ultimately, the success of STARburst underscores the power of user-centric support in building a self-sustaining R ecosystem, poised to meet GSK's ambitious goals and embrace the future of pharmaceutical analytics.

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. {gt} package changelog: <https://gt.rstudio.com/news/index.html#gt-110>
11. CAMIS main site: <https://psiaims.github.io/CAMIS/>

CONTACT INFORMATION

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

Author Name: Benjamin Arancibia
Company: GSK
Email: benjamin.c.arancibia@gsk.com

Author Name: Becca Krouse
Company: GSK
Email: becca.z.krouse@gsk.com

Brand and product names are trademarks of their respective companies.