

SCE Roundtables

A forum for Scientific Computing Environment product owners

R/Pharma contributors

This document captures discussions across R/Pharma in 2023

Table of contents

Scope and purpose	5
Definitions	6
1 Modern SCEs	7
2 Topics discussed	8
2.1 Is the next step Homegrown or vendor	8
2.1.1 Relationship with informatics partners	8
2.2 What is an SCE?	8
2.3 Building trust in business/open source code	9
2.4 Change-management into a modern SCE	9
2.5 General notes	10
3 Actions	11
4 Validate shiny?	12
5 Question	13
6 Topics discussed	14
6.1 Do we need to validate?	14
6.2 Robust UIs?	14
7 Actions	15
8 Change management	16
9 Question	17
10 Who are our people?	18
11 Theme: Emerging Talent	20
12 Theme: Other Points and Considerations	21
13 Multi-modal drug development	22

14 Question	23
15 Topics discussed	24
16 Depending on OS	25
17 Question	26
18 Interactive CSR	27
19 Question	28
20 The case for OS	29
21 Question	30
22 Contributors	31
22.1 Round table advisory board	31
22.2 Participants	31
22.3 Organising committee	32
22.4 Advisory board	32
References	34

Scope and purpose

This document captures discussion, pain points and a path to next steps after the R/Pharma events of 2023. There are two events that led to this document being created.

F2F Roundtables in Chicago

~60 leaders from 40+ companies met F2F in Chicago for a series of discussions on the most pressing topics for late-stage reporting in R.

The discussion was crowdsourced via a github discussion, and led to the following topics:

- [What are our goals for a modern clinical reporting workflow, on a modern SCE?](#)
- [What are the risks with our increasing external business code dependencies?](#)
- [We have a path to R package validation - but what are we doing with shiny apps?](#)
- What is the path to an interactive CSR?
- The case for contributing to OS
- Where are we with our people?
- What should we be doing to leverage advances in LLMs/AA/AI impact? (at the drug development through to developer efficiency levels)
- What are the barriers bringing imaging/genomics/digital biomarkers and the CRF closer?

References

[Promotional site for round-tables R Validation Hub update Doug's slides on the shared validated repo CAMIS - comparing differences based on tool used](#)

Definitions

Abbreviation	Description
SCE	Statistical/Scientific Computing Environment

1 Modern SCEs

What are our goals for a modern clinical reporting workflow, on a modern SCE? What are our learnings today achieving that goal, and how can we better prepare ourselves to balance the drive to innovate while having to evolve people and processes?

Chairs: James Black and Satish Murphy

2 Topics discussed

2.1 Is the next step Homegrown or vendor

- Split in vendor approaches (GSK, AZ) vs homegrown (Roche, J&J, Amgen) for the new generation. We don't know what the ideal is today, but we know we need to be able to evolve and adapt to new technologies and approaches much more than we did in the past.
- Homegrown usually means modular, as still reliant on different open source and vendor solutions (e.g. AWS, Hashicorp, etc). Is a more accurate description turnkey vs modular?
- How to ensure our modular platform scales is an important new aspect, especially using new open source tools.
- In a turnkey, the vendor will have baked in provenance. In a modular we must focus on making sure metadata/provenance runs as a background across the system to ensure we can trace back from an insights and transformations to the source data.
- A major pain point is how things are funded - we are not used to funding a platform for sustained evolution/innovation, but in the data science space things are constantly evolving - so moving to operations/maintenance is equivalent to decay.

2.1.1 Relationship with informatics partners

- There is often tension between informatics and the business, with the growth of business written code that looks more like software (e.g. R packages) vs the the scripts/macros we use to make. We shared experiences finding a balance as we entered this new phase.

2.2 What is an SCE?

- Should GxP and exploratory remain separate platforms?
 - Split across companies in the group, with some companies having a single platform for both, and others having separate platforms.
 - With initiatives like the digital protocol coming, we don't know what the impact will be on routine clinical reporting (but could become further optimised)
 - Pain points merging:

- * validation (CSV) is a long process in most companies, which can impact ability to support exploratory work.
- * Needs are different. E.g. clinical reporting is low compute, while design and biomarker work is often heavy in memory and data.
- Is data part of the SCE? Traditionally yes, but some but not all companies are decoupling data from compute.
- Whether it's in the SCE or not, traceability is uniquely important in our domain of regulatory reporting.
- It appeared across all companies access to an SCE is now through a web-browser (not a local application)

2.3 Building trust in business/open source code

- [The Cathedral and the Bazaar by Eric Rayman](#) was a recommended essay to read, that talks about 'Cathedral' products where the code is developed in a closed environment then released, vs 'Bazaar' products where the code is developed in the open. An argument is the Bazaar model, as long as it is a project with enough eyeballs, will lead to shallow bugs; this is also known as [Linus' Law](#).

2.4 Change-management into a modern SCE

- What are we actually building? A general data science platform? A platform optimised for clinical reporting?
 - These are not the same platform, and which you pick has an impact. e.g. should statisticals programmers learn git, or should we give a simple GUI for pushing code through QC and to Prod?
 - There is not a consensus about this for next-gen, with only a handful of companies expecting statistical programmers to work in the same way general data scientists.
- Historically we depended on SAS, it's data formats, and filesystems. How to build a modern SCE that doesn't?
 - Do we enable legacy workflows to work in the new SCE? Only new ways, or how do we find a balance to ensure business continuity while enabling innovation?
 - The human and process change management piece is massive, and SCE POs must work in tandem with statistical programming leadership.
 - Agreement the biggest pain point is the dependency on filebased network share drives for data and insight outputs. One company mentioned they have millions of directories in their legacy SCE.

- Most companies have carried over having the outputs server be a network share drive, but would a more ‘publishing’ type model be more robust?

2.5 General notes

- We manage on user access. A question is whether we can control access based on user access, and the intended use. In terms of both where they are working and what the context of the work is.
- We need to rightsize our ambitions, as going to broad will slow us down.
- How will moving to this latest generation be a positive impact on our financials? Interesting discussion putting ourselves in a CMO’s shoes - if you don’t care about how the CSR is generated, how is the new SCE making the company money?
- Interactive analysis is growing - need to prepare for when people want to use something like shiny for GxP
- The ideal people to work on the SCEs are unicorns - they need to be able to work with the business, understand the trial processes, and be able to work with the technology. We need to be able to train people to be unicorns, and we need to be able to retain them.

3 Actions



- Can we have a learning series / discussion / panel on experience using a designing git flows
- What are companies experiences across the diverse scopes we've placed onto SCEs - need to share more here.
- How do we want to store data post shared drive, how do we want to track?
- We need to share learnings on the value add from these new SCEs!

4 Validate shiny?

Chairs: James Black and Harvey Lieberman

5 Question

We have a path to R package validation - but what about shiny apps? In what context would validation become relevant to shiny app code, and how can we get ahead of this topic to pave a way forward for interactive CSRs?

6 Topics discussed

6.1 Do we need to validate?

- Tiered approach / decision tree
 - Lowest is made by study team for study team. 2nd level is risk is unsupervised use, or specific contexts - e.g. making an app for dosing or safety. 3rd would be shiny CSR.
 - Is the results going directly from the app into a submission?
 - Don't validate a shiny app - validate the static functions in the R packages. CSV may not be relevant for UIs (vs static R packages)

6.2 Robust UIs?

- Good to have unit tests - often manual testing. Automated can easily get messed up as the code evolves.
- We should use the git flow - e.g. protect master and disable manual deployments
- Show or download R code is perfect for reproducibility → e.g. show code button
 - But then need to actually run that in a prod batch run
 - this use can case skip validation as code is run as study code
- Some cases where you don't want to export and run code → e.g. output used directly for decision making are coming
- How to handle risk of UI problems if our focus is on the static code - e.g. misnamed reactive values so wrong values being shown, even if static R packages giving correct results.
- Risk based is really important - e.g. for something like dark mode breaking, we need to know what requirements are high risk (e.g. table is correct) vs low risk (e.g. dark mode button)

7 Actions



- Can we share some common high level guidance on stratifying risk in shiny shared across companies? (Pfizer has written this already internally).
- Discuss if we should have an extension of R package whitepaper to cover shiny?

8 Change management

Chairs: Matthew Kumar and Cassie Burns

9 Question

Where are we on getting data analysts and data scientists that work with clinical data on board (in particular, those delivering CSRs and submission packages)? What are the challenges - what has been overcome?

10 Who are our people?

Prefaced both sessions by asking individuals to define the *our* in *our people*;

- Stat Programmers
- Statisticians
- Data Management
- Other CSR-deliverable oriented roles (e.g. medical and scientific writing)
- Management, Leadership

Theme: R Adoption and Challenges

- The adoption of R requires varied types of commitment depending on the perspective of the stakeholders involved, notably management and employees.
- Leadership usually supports the adoption of R, yet, in many cases, they don't adequately communicate or advocate its application. Common concerns include the lack of realized ROI and the perception of R as a "nice-to-have" rather than a necessity.
- It is not viable to mandate or compel individuals to learn R.
- Seasoned programmers, who prefer proprietary software, may leave the company if forced to switch.
- These programmers often prefer to maintain current workflows that involve proprietary tools, established macros, homegrown IDEs, etc.
- Some in management endorse an approach of "mandate" or "force," while others aim to "encourage."
- Experienced stat programmers cite R's learning curve as an obstacle to transition, and some don't see the ROI in making the switch.

Theme: Change Management

- Implementing proper change management was emphasized by several attendees in both sessions.
- Organically, through change management, approximately 25% of experienced statistical programmers or statisticians have successfully completed R training, while the remaining 75% have shown resistance.

- Among the successful 25%, only 5% have applied what they've learned in actual study work. This is often due to time constraints related to product deliverables.
- Mapping the transition to R with learning and development goals is one strategy.
- A structured learning plan and a roadmap for R upskilling are essential. This includes trainings focused on R, particularly in the context of the pharmaceutical industry, and from a proprietary software programmer's perspective.
- Identification of champions or early adopters among statistical programmers could aid in transitioning colleagues.
- Several companies shared their strategies for promoting community learning (e.g., bi-monthly meetings, presentations, assignments), both on a "just in time" basis and on a regular schedule.
- Pointing individuals to ongoing efforts and resources, such as R in Pharma, PharmaVerse, Phuse, etc., can boost awareness and participation.
- Granting individuals protected or dedicated time to learn and fail is recommended. An analogy used was "giving them a safe sandbox to try making a castle."
- R need not be used for all tasks immediately. A more measured approach, such as starting with creating figures and then moving to more complex tasks, like AdAM programming, could better build confidence and competence.
- Ensuring enough transition time and clear direction ("as of X date, we'll work in R") is crucial.
- Having leadership advocacy is vital at the end of the day.

11 Theme: Emerging Talent

- Newer talent is increasingly trained in open-source approaches and languages, with fewer exposed to proprietary tools.
- With the rise of data science as a field of study, many are less interested in joining a company for routine implementation work; they identify as “data scientists” and have been trained in markedly different ways.
- This affects talent attraction, development, and retention within a company.
- Innovation can come from new hires, justifying the need to foster their development and listen to their insights.
- There’s a unique opportunity for co-mentorship: new hires (proficient in R, Python, etc.) and existing staff (experts in domain knowledge, clinical trials, etc.): *how vs what/why*
- There’s a need for clear “career pathing” or “trajectories” within statistical programming as roles evolve. Examples include:
 - “Analyst” requires traditional statistical programming knowledge and training
 - “Engineer” needs DevOps skills and a systems mindset
 - “Tool builder” needs a software engineering mindset
- General trends suggest companies are demanding a secondary language in addition to proprietary software (not necessarily R), but knowledge of at least two languages indicates an individual could reasonably learn R.

12 Theme: Other Points and Considerations

- Questions to consider include: *What kind of training will people need in the future state? How should the support be arranged to enable the future state, potentially with IT and DEV involvement?*
- Due to the required skillset, statistics and programming now need to work together more than ever
- Stakeholders seek the benefits of R (e.g., Shiny, Rmarkdown), but often lack personnel to build and maintain these assets.
- R and Shiny tools can be utilized in more areas beyond TLF programming such as dose decision meetings, clinical trial design, administrative tasks, and long-term-focused applications.
- Legacy infrastructure (e.g., virtual machines and proprietary software) can pose challenges when implementing newer approaches like R and Shiny, making the transition difficult and cumbersome.
- AI and GPT can be a valuable tool in transitioning to R, but won't completely replace a programmer. It can be used to effectively explain or translate existing code or generate entirely new code.

13 Multi-modal drug development

Chair: Katie Igartua

14 Question

There is more need than ever to integrate different roles, and ways of working, along with different data modalities. What are the barriers bringing imaging/genomics/digital biomarkers and the CRF closer, how could we overcome them, and what is our envisioned benefit?

15 Topics discussed


1. Use of real-world evidence data (RWE) for contextualizing clinical trial samples to support indication selection, patient settings and combination therapy strategies.
 - Challenges for users arise when leveraging multiple sources (both public and licensed) given biases such as in abstraction rules or genomic assays.
 - Best practices of real world evidence outcomes analyses (eg. rwPFS, rwOS).
2. Integration of Claims datasets and validation. Requirement for multiple lines of evidence for a given event would enrich the quality and usability of the data and bypass biases from the source of claims data.
3. Imaging validation frameworks. Challenges discussed include i) interpretability and adoption of deep networks models and utility relative to the gold standard (e.g. prediction vs. RESIST criteria), ii) transferability of models across different instrument platforms and iii) variability of pathologist vs. radiologist calls in the labels.
4. Use of smart devices in clinical trials. Consensus was that this is more common in non-oncology areas (e.g. cardio). How can we mitigate risk of compliance in trials?
5. Contextualizing small patient cohorts with rich phenotype data and longitudinal data. Liquid assays for monitoring resistance mechanisms in oncology.

16 Depending on OS

Chairs: Mike Smith & Ed Lauzier

17 Question

How much risk is there in depending on external packages, and can we foster a clearer set of expectations between developers and people/companies that depend on these packages?

 Missing notes


Content is still coming, an email will be shared once the site is complete.

18 Interactive CSR

Chairs: Ning Leng and Phil Bowsher

19 Question

If we assume it's technically possible to transfer - what would it mean to give an interactive CSR? How would primary, secondary and ad-hoc analysis be viewed? Would views change depending on role (e.g. sponsor, statistical reviewer, clinical reviewer)?

 Missing notes


Content is still coming, an email will be shared once the site is complete.

20 The case for OS

Chairs: Satish Murphy & Becca Krouse

21 Question

What is stopping people and companies from contributions in OS? Can we define the case for contributing to OS?

 Missing notes

Content is still coming, an email will be shared once the site is complete.

22 Contributors

22.1 Round table advisory board

Ordered alphabetically by company;

- Cassie Milmont; Amgen
- Lee Min; Amgen
- Michael Blanks, R/Pharma executive; Beigene
- Ning Leng, R/Pharma organizing committee; Genentech
- Doug Kehlkoff, R Validation Hub Lead; Genentech
- Michael Rimler; GSK
- Andy Nicholls; GSK
- Volha Tryputsen, R/Pharma organizing committee; J&J
- Sumesh Kalappurakal; J&J
- Mark Bynens; J&J
- Harvey Lieberman, R/Pharma executive; Novartis
- Shannon Pileggi; The Prostate Cancer Clinical Trials Consortium (PCCTC)
- Mike Smith; Pfizer
- Max Kuhn; Posit
- Rich Ioannone; Posit
- Paulo Bargo, R/Pharma executive
- James Black, R/Pharma executive; Roche

22.2 Participants

Ordered alphabetically by company;

- Rose (Abbot Labs)
- Mike (Astellas)
- Limin (Atmos)
- Daniel (Astrazeneca)
- Kevin (BI)
- Mathias (BI)
- Eric (Biogen)

- Mary (BMS)
- Nicole (Denali)
- Mike Thomas (Flatiron Health)
- Daniel (Formycin)
- Satish Murphy (Johnson & Johnson)
- Nick (Johnson & Johnson)
- Jan (Moffat Cancer Centre)
- Harvey (Novartis)
- Li (Onc)
- James Kim (Pfizer)
- Mike (Pfizer)
- Michael Meyer (Posit)
- Doug (Roche)
- Jeeva (Roche)
- James Black (Roche)
- Andrew (Sanofi)
- Abigail (Tempest)
- Susheel (Vertex)
- Derek (WL Gore)

22.3 Organising committee

- Phil Bowsher, R/Pharma executive; Posit
- James Black, R/Pharma executive; Roche
- Harvey Lieberman, R/Pharma executive; Novartis

22.4 Advisory board

- Cassie Milmont; Amgen
- Lee Min; Amgen
- Michael Blanks, R/Pharma executive; Beigene
- Ning Leng, R/Pharma organizing committee; Genentech
- Doug Kehlhoff, R Validation Hub Lead; Genentech
- Michael Rimler; GSK
- Andy Nicholls; GSK
- Volha Tryputsen, R/Pharma organizing committee; J&J
- Sumesh Kalappurakal; J&J
- Mark Bynens; J&J
- Harvey Lieberman, R/Pharma executive; Novartis
- Shannon Pileggi; The Prostate Cancer Clinical Trials Consortium (PCCTC)

- Mike Smith; Pfizer
- Max Kuhn; Posit
- Rich Ioannone; Posit
- Paulo Bargo, R/Pharma executive
- James Black, R/Pharma executive; Roche

References