Open Source Technology in Clinical Data Analysis

PHUSE

December 4, 2023

A significant amount of time and energy has been invested in recent years exploring the desirability (do we want it?), feasibility (can we do it?), and viability (is it worth it?) of integrating open source solutions into our clinical data pipelines which transform source data into clinical study reports and submission data packages. When this manuscsript is complete, we hope to put to rest some of the burning questions that we believe we now know the answers to. This will allow industry, and all the passionate people in it, to look ahead and start tackling the next horizon of challenges related to using open source solutions for clinical data pipelines. We hope you will contribute your expertise to this effort.

Table of contents

Preface **6**
DRAFT (December 2023) 6

1 OSTCDA **7**
1.1 Purpose and Background 7

2 What is Open Source? **8**

3 What is Open Source? **9**
3.1 How to Contribute 9
3.2 Guidance 9
3.3 Examples 9

4 Why Open Source? **10**
4.1 What is the 'why' for using open source solutions in pharma clinical data analytics? 10
4.2 How to Contribute 10
4.3 Guidance 10

5 Establishing Trust **11**
5.1 Can an open source solution be trusted? 11
5.2 How to Contribute 11
5.3 Guidance 11

6 Documenting Trust **12**
6.1 How do you document your trust in an open source solution? 12
6.2 How to Contribute 12
6.3 Guidance 12

7 Cost of Open Source **13**
7.1 What is the true cost of implementing open source solutions? 13
7.2 How to Contribute 13
7.3 Guidance 13

8 Regulatory Acceptance **14**
8.1 Will the regulatory agencies accept data and analyses generated with solutions developed and available as open source? 14

8.2 How to Contribute . . . 14
8.3 Guidance . . . 14

9 GxP Compliance **15**
9.1 How do you establish reproducibility and traceability? . . . 15
9.2 How to Contribute . . . 15
9.3 Guidance . . . 15

10 User Support **16**
10.1 How do we support users in managing an ever-evolving environment of Open Source solutions? . . . 16
10.2 How to Contribute . . . 16
10.3 Guidance . . . 16

11 User Development **17**
11.1 How do you transform the traditional Statistical Programmer into the future Data Scientist? . . . 17
11.2 How to Contribute . . . 17
11.3 Guidance . . . 18

12 Numerical Matching **19**
12.1 Do we need to match SAS numerically when using a different language? . . . 19
12.2 How to Contribute . . . 19
12.3 Guidance . . . 19

13 OS in the Long Run **20**
13.1 How do we ensure that the solutions being developed today will exist in the long run? . . . 20
13.2 How to Contribute . . . 20
13.3 Guidance . . . 20

14 Funding OS **21**
14.1 Is it possible for industry fund open source? . . . 21
14.2 How to Contribute . . . 21
14.3 Guidance . . . 21

15 Liability with OS **22**
15.1 Are contributors to open source exposing themselves to any liability of their solutions? . . . 22
15.2 How to Contribute . . . 22
15.3 Guidance . . . 22

16 Legal Concerns **23**
16.1 Are there any legal concerns from open source development? . . . 23

16.2 How to Contribute 23
16.3 Guidance 23

17 OS Business Models **24**
17.1 What open source models are available for businesses? 24
17.2 How to Contribute 24
17.3 Guidance 24

18 What Else? **25**
18.1 What else can we do? 25
18.2 How to Contribute 25
18.3 Guidance 25

References **26**

Preface

DRAFT (December 2023)

The information contained in this Quarto book we aim to comprehensively address the most important questions related to deploying open source solutions for clinical data analytics in the pharmaceutical and vaccine development industry.

- What questions have been asked and ***already*** answered?
- What questions have been asked and ***nearly*** answered?
- What questions have been asked and ***not yet*** answered?

We are developing this manuscript in the open and accepting contributions by the community via our GitHub repository's Discussions tab. Please contribute your thoughts, perspectives, references, citations, and links through that mechanism. We'd like to be able to attribute your ideas to you, so providing the rationale supporting your thoughts will strengthen your argument. Please be as thoughtful and thorough in your contributions as you can! You can also upvote questions and/or responses that you find particularly valuable.

At the moment, this document is in draft form, hence please ***DO NOT CITE*** it as a reference.

To learn more about this initiative, please watch our **[2023 R in Pharma]**(https://rinpharma.com/){target="_ lightning talk The State of Open Source Technology in Clinical Data Analysis, Reporting, and Submissions.

Thank you to PHUSE for supporting this endeavor. And thank you to ***YOU*** for your thoughtful contributions to the effort.

1 OSTCDA

Open Source Technology in Clinical Data Analysis

A significant amount of time and energy has been invested in recent years exploring the desirability (do we want it?), feasibility (can we do it?), and viability (is it worth it?) of integrating open source solutions into our clinical data pipelines which transform source data into clinical study reports and submission data packages.

This repository will serve to collect and synthesize expert opinions and resources for a (hopefully) comprehensive set of questions which arise as organizations travel this journey.

The discussions will provide citable (and sometimes quotable) input from industry experts, resulting in a "state of the union"-style manuscript that will help us move past questions that have already been sufficiently addressed and focus on those that remain.

We invite you to navigate to the Discussions section section to provide your thoughts, resources, or perspectives that help address any or all of the questions. If we've overlooked a key question - start up a new discussion thread!

1.1 Purpose and Background

There are many questions around understanding and using open source for clinical data analysis. We want to create a comprehesive knowledge base about the "state of the union" and provide an overview and ideally also answers for core questions. We need and collect input from our community to compile the knowledge base, so please join the discussions to allow a broad and complete picutre.

If you like to know more, please join the R/Pharma talk "The State of Open Source Technology in Clinical Data Analysis, Reporting, and Submissions". The recording is here: 2023 R/Pharma presentation.

2 What is Open Source?

3 What is Open Source?

When someone says ‘open source’ - what does that mean to you?
What are the important characteristics of something that is regarded as ‘open source’?

- Does the cost matter?
- Does the ability to review the code matter?
- Does the ability to reuse the code matter?
- Does the ability to modify the code matter?

3.1 How to Contribute

Contribute to the discussion here in GitHub Discussions:
What is open source?

3.2 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

3.3 Examples

Example 1: The Open Source Initiative provides a definition:

- Free distribution
- Available source code
- Derived works possible

Example 2: Merriam-Webster defines open source as “having the source code freely available for possible modification and redistribution”.

4 Why Open Source?

4.1 What is the 'why' for using open source solutions in pharma clinical data analytics?

- What is the attraction to open source solutions?
- Why do users like open source solutions?
- Why are leaders of organizations in Data Management, Biostatistics, and Programming devoting resources toward the development, testing, and adoption of open source solutions?

4.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
What is the 'why' for open source?

4.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

5 Establishing Trust

5.1 Can an open source solution be trusted?

- How do we have confidence that an open source solution is accurate?
- What are the relevant considerations?

5.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
Can an open source solution be trusted to be accurate?

5.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

6 Documenting Trust

6.1 How do you document your trust in an open source solution?

- How do we have document our trust that an open source solution is accurate?
- How do we know if a third-party will accept our documentation of trust?

6.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
How do you document your trust in an open source solution to satisfy a third-party inquiry?

6.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

7 Cost of Open Source

7.1 What is the true cost of implementing open source solutions?

- Is it essentially free?
- What resources are required for proper implementation?

7.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
What is the true cost of implementing open source solutions into clinical data analytic processes??

7.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

8 Regulatory Acceptance

8.1 Will the regulatory agencies accept data and analyses generated with solutions developed and available as open source?

- What do we know regarding data submissions to FDA?
- What do we know regarding data submissions to other regulatory agencies?
- Are there technical considerations for the creation of submission data packages?

8.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:

1. Will the **FDA** accept data and analyses generated with solutions developed and available as open source?
2. Will **other regulatory agencies** accept data and analyses generated with solutions developed and available as open source?

8.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

9 GxP Compliance

9.1 How do you establish reproducibility and traceability?

GxP compliance means establishing accuracy, reproducibility, and traceability. When working with open source solutions to process and analyze clinical trial data:

- How do we establish reproducibility of the outputs?
- How do we establish traceability of the input through to the output?

9.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
How do you establish reproducibility and traceability with open source solutions?

9.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

10 User Support

10.1 How do we support users in managing an ever-evolving environment of Open Source solutions?

The conventional user (programmer) processing clinical trial data may be used to stability of the available toolbox at their disposal.

- How can we help users to operate in a (potentially) more variable environment?
- How can we help users to address unexpected challenges due to changes in their computational environment?

10.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
How do we support users in managing an ever-evolving environment of Open Source solutions?

10.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

11 User Development

11.1 How do you transform the traditional Statistical Programmer into the future Data Scientist?

The traditional Statistical Programmer/Analyst in pharmaceutical and vaccine development primarily fulfills their role using the SAS programming language to develop single-use scripts that read in data and create an output dataset or analysis display.

Many Data Scientists are:

- programmatically multilingual
- leverage open source tools
- are familiar with object-oriented languages
- develop code collaboratively with platforms such as GitHub
- version control code with technologies such as git
- are comfortable having code reviewed for functionality and good programming practice
- are familiar with good software development practices
- are familiar and comfortable with agile ways of working

How will we transform the traditional Statistical Programmer into the future Data Scientist?

11.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
How do you transform the traditional Statistical Programmer into the future Data Scientist?

11.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

12 Numerical Matching

12.1 Do we need to match SAS numerically when using a different language?

- What if we the same inputs yield similar, but numerically different results?
- What if we the same inputs yield drastically different results?
- What is the truth? Which is correct?
- What if SAS and R are equivalent, but a third language yields numerical differences?

12.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
Do we need to match SAS numerically when using a different language?

12.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

13 OS in the Long Run

13.1 How do we ensure that the solutions being developed today will exist in the long run?

When these solutions are embedded in data pipelines, if development and maintenance support disappears, there is a risk to the pipelines which leverage them.

- How do we ensure long term viability?
- How do we ensure long term sustainability?
- How do we ensure long term maintainability?

13.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
How do we ensure that the solutions being developed today will exist in the long run?

13.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

14 Funding OS

14.1 Is it possible for industry fund open source?

- Is it possible to fund OS development in pharmaceutical drug and vaccine development?
- What might a funding model look like?
- What problem(s) would funding solve?
- Are there examples of this in other industries?

14.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
Is it possible for industry fund open source?

14.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

15 Liability with OS

15.1 Are contributors to open source exposing themselves to any liability of their solutions?

- What are possible sources of liabilities?
- Are there mitigating actions which can limit or eliminate liabilities?

15.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
Are contributors to open source exposing themselves to any liability of their solutions?

15.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

16 Legal Concerns

16.1 Are there any legal concerns from open source development?

- What do individuals need to know?
- What do organizations need to know?
- How does this differ if the solution is an individual or a collaborative effort?

16.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
Are there any legal concerns or ramifications from open source development (on the user, developer, organization)?

16.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

17 OS Business Models

17.1 What open source models are available for businesses?

- How to address the challenge of making money providing software that is, by definition, licensed free of charge?
- What is open core?
- Wikipedia page

17.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
What open source models are available for businesses?

17.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

18 What Else?

18.1 What else can we do?

This is very open ended and initially intended as a catch-all for the future considerations not already covered by the other discussions.

- What have we missed?
- What other questions are being asked and/or need to be addressed?

18.2 How to Contribute

Contribute to the discussion here in GitHub Discussions:
What else can we do?

18.3 Guidance

- Provide your thoughts and perspectives
- Provide references to articles, webinars, presentations (citations, links)
- Be respectful in this community

References