# Open Source Technology in Clinical Data Analysis

PHUSE

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.

# Table of contents

# Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

# 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

# References