COSA - SDTM Open Sourcing Proposals

OAK Garden - SDTM Automation
The flourishing Data Transformation Engine
Dec-2022

OAK - Introduction

Algorithms - Deep Dive

raw.synthetic.data - R package

CDISC COSA Proposals

1. {oak} as a open-source
2. PoC to automate SDTM based on CDASH standards
3. {raw.synthetic.data} a open source solution

Next Steps

OAK Garden SDTM Automation

OAK Garden is part of the "next-generation" solution for Roche's data and analytics platforms to move towards increased automation.

The OAK Garden will be used by the Data Science teams to produce SDTM.

Contributes to the prospective FAIRification of clinical trial data by creating SDTM datasets integrated with the Global Data Standards to ensure interoperability of the data.

PDD Next Generation Tools

For regulatory reporting

Driven by metadata & Global Data Standards, OAK Garden can

Automate ~80% SDTM domain with

~22 Reusable Algorithms.

OAK Garden - Benefits

Time Saving

Allows to generate SDTM deliverables **at least 50% faster** than with the current process based on adherence to Global Data Standards.

Automation

- Innovative solution
- Automates the most tedious work (e.g. SDTM Specifications)

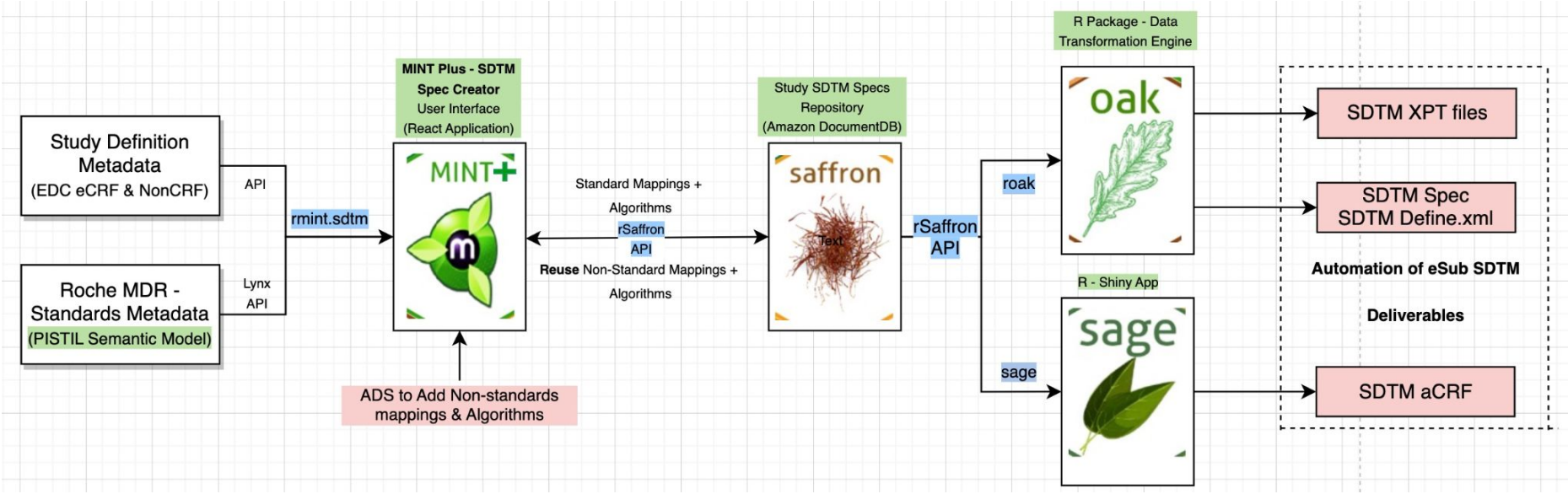
Simplicity

- Everything **bundled into 1** package
- Suited for **beginner** R programmers

OAK Garden Components

Pistil & HoneyBee (Semantic Model)	Components of Roche MDR. A Graph that stores raw source to target SDTM mappings in a machine readable format. This also hosts the metadata(algorithms) required for automation
MINT+ (SDTM Spec creator)	React Web Application to create study SDTM mappings which automates the standard mappings in the study based on Roche MDR. Enables adding Non-standard mappings in study.
Saffron (Study SDTM spec Repository)	Stores the study SDTM spec in a machine readable format (JSON). Enables reuse of Non-standards SDTM mappings across studies,
oak - Data Transformation Engine	An R package that drives the automation of SDTM using metadata.

OAK Garden - Metadata Flow



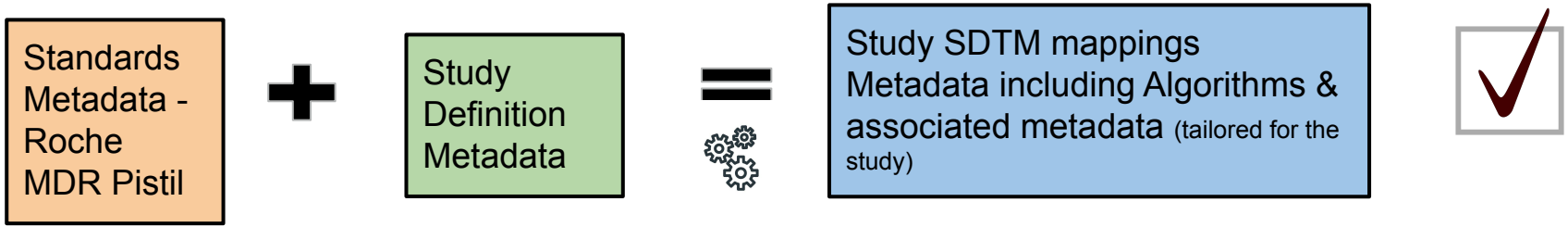
What is Metadata?

Metadata is "data about data".

For example, for a Clinical study, the Study Definition Metadata is

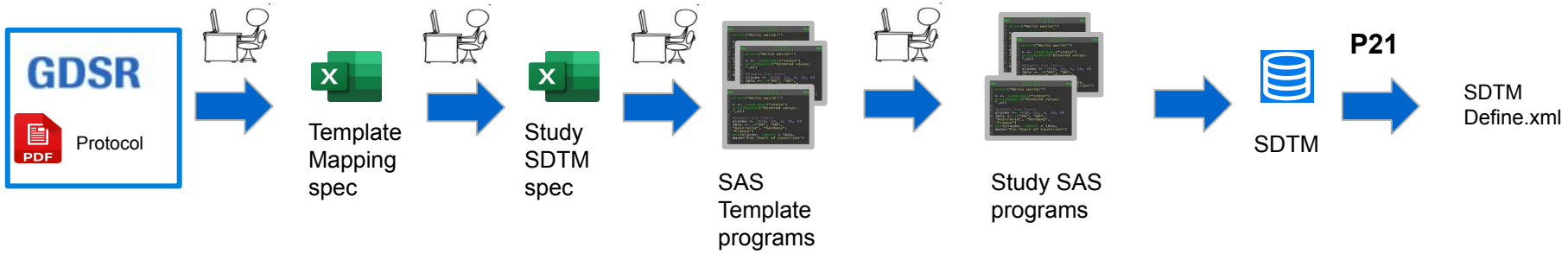
CRF - EDC or ODM
- FormOID (RAW dataset name)
- FieldOID (RAW variable name)
- Data Dictionary (Study Codelists)

Non-CRF - Vendor specification Definition
- Dataset Name (RAW dataset name)
- Variable Name (RAW variable name)
- Appendix from FFS (Study codelists)



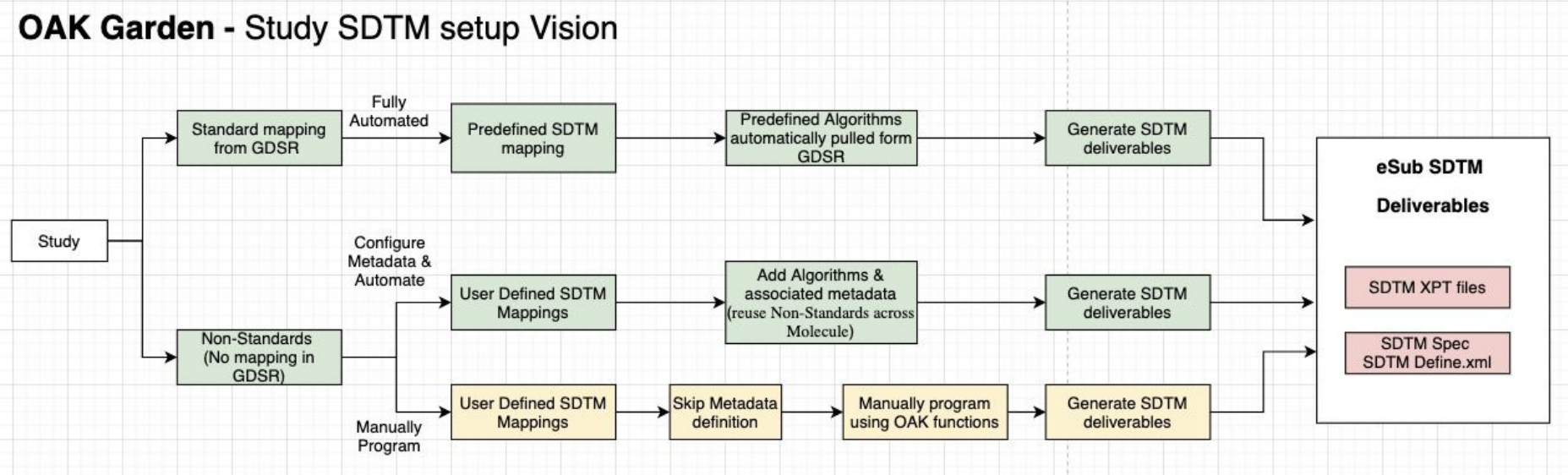
OAK Garden - Metadata driven Automation

Current Workflow: Manually implement SAS programs & specs



Future Workflow: Automated SDTM Datasets & specifications.

OAK Garden - Study SDTM setup Vision



★ **Automation of Standards -** Closely linked to Roche MDR, standards are automated out of the box. Based on the studies started after 2019, a study uses 82% (median) Data-standards. This means, we can expect on an average of 80% SDTMv automation for every study when using OAK Garden.

★ **Flexibility to Automate Non-Standards** - Driven by the ADS, MINT+ UI & Saffron enables storing and reusing the Non-standard SDTM mappings & Algorithms across studies. ADS can browse Non-standards already existing in Saffron using the Intuitive MINT+ User Interface and use it in their studies along with the previously used Algorithms.

★ **Manual Programming -** Flexible Architecture, enables ADS to program Non-Standards for complex scenarios in R or in SAS.

Algorithms - Deep Dive

Algorithms - *Core Concept*

CORE CONCEPT: **REUSABLE ALGORITHMS**

- SDTM Mappings are defined as algorithms that transforms the collected (CRF, nonCRF) source data into the target Tabulation data model. Mapping Algorithms are the back bone to the SDTM automation.

We have designed 22 unique mapping Algorithms to accommodate most (80%) of the TA standards & Non-CRF data models

22 Unique Algorithms

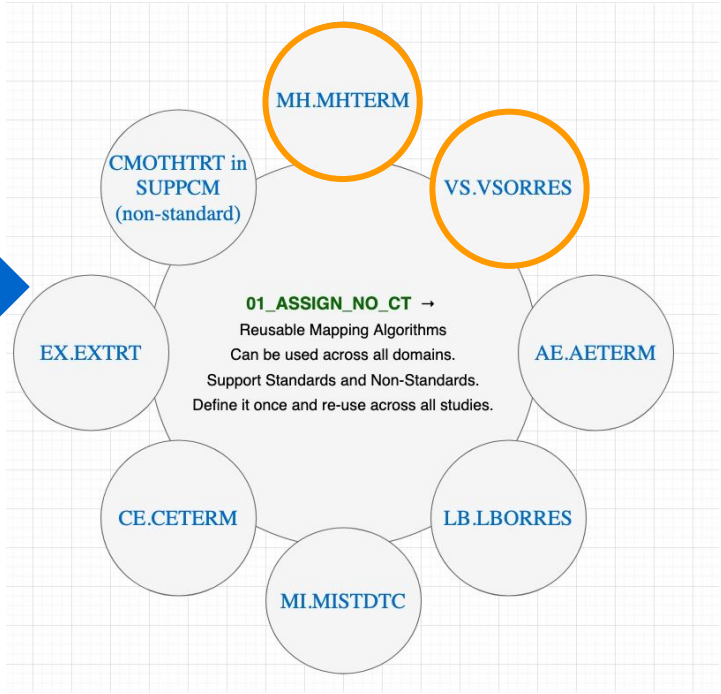
Key Points:

- Algorithms can be re-used across Domains
- Algorithms can be **pre-specified** for Standards
- Users can reuse/add algorithms for non-standards or new data types
  - **Both Standards and Non-Standards can therefore be supported**
- Programming language agnostic - this concept do not rely on a specific programming language for implementation. We have implemented them as R functions.

Reusable Algorithms - Example

The algorithms can be applied in many different contexts (see right)

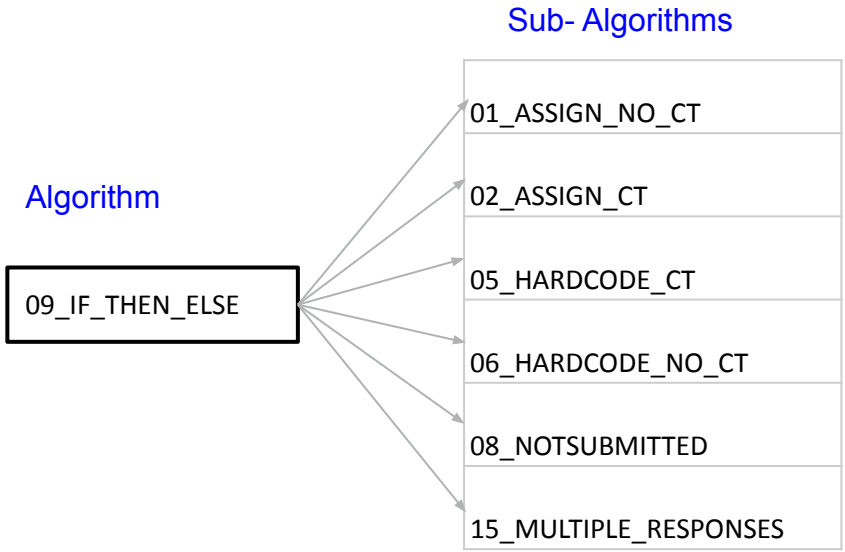
Mapping Algorithm	Description
01_ASSIGN_NO_CT	One to One mapping with no controlled terms.
02_ASSIGN_CT_ST	One to One mapping with controlled terms.
05_HARDCODE_CT	Hard code the target based on the source with controlled terms.
06_HARDCODE_NO_CT	Hard code the target based on the source without controlled terminology.
09_IF_THEN_ELSE	Conditionally check a condition and apply a mapping

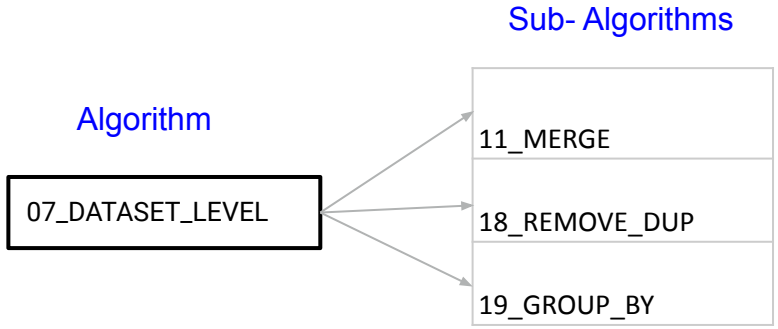


Algorithms & Sub-Algorithms

Only as Algorithms	Only as Sub-Algorithms	Algorithms & Sub-Algorithms
03_AE_AEREL	11_MERGE	01_ASSIGN_NO_CT
07_DATASET_LEVEL	18_REMOVE_DUP	02_ASSIGN_CT
09_IF_THEN_ELSE	19_GROUP_BY	05_HARDCODE_CT
17_WHODRUG_FA	20_NEED_USER_INPUT	06_HARDCODE_NO_CT
13_RELREC		08_NOTSUBMITTED
14_RELREC_CONDITION		15_MULTIPLE_RESPONSES
21_NONCRF_LAB		
22_NONCRF_PKC		
23_PAIRED_VARS		

Algorithms & Sub-Algorithms





The permutation & combination of Algorithms & sub-Algorithms creates endless possibilities to accommodate different types of mappings.

Demo!

raw.synthetic.data - R package

raw.synthetic.data - Current State Challenges

Reliance on vendor test data (timeline impact, data quality)

Test data entered manually to the EDC which is used for SDTM and other clinical programming activities.

Limitation and cumbersomeness of current process (i.e. create test study in Rave & enter data manually)

Challenges shared across but currently - company specific solutions or lack of solution.

Time consuming. Often in a critical path limiting the programming and QC of SDTM & other clinical programming activities.

Manually entered test data is not accurate or Biologically correct.

Raw Synthetic Data Project

- Create raw synthetic data based on the study design for EDC systems and nonCRF/vendor data.
- Use advanced analytics and create “Biologically correct” synthetic data.
- Automate & accelerate SDTM programming or any clinical programming tasks at the study start
- Collaborate with other companies to create this automated solution that is EDC agnostic with an aim to open source so that any pharma company can utilize it or contribute to the development and maintenance of the package
- Metadata / schema driven approach & may need publicly available data

raw.synthetic.data R package - Roche version

Metadata driven & EDC/Vendor agnostic framework to generate synthetic data

Limits the test data creation to Data Dictionaries (aka codelists) when available.

Create “some” test data for Numeric & Character text boxes.

Adheres to the Primary Matrix (Visit-Form schedule) defined in study.

Create test data across multiple visits/sites.

Generates test data for EDC system variables.

Directly grab from global standards libraries as we develop new standards and study metadata file to create synthetic data

Enables team automate and accelerate programming tasks (i.e. raw synthetic data > synthetic SDTM > synthetic ADaM which is continuously aligned with global standards

Demo!

COSA Proposals

1. {oak} as a open-source
2. PoC to automate SDTM based on CDASH standards
3. {raw.synthetic.data} a open source solution

COSA - Proposal 1 - {oak} as a open-source

Vision

- Open source modularized toolbox that enables data scientists to develop SDTM datasets in R.
- Unlike Roche version, this is not an automation solution, instead this package has useful functions to program SDTM in R.
- Follow ODM standards & remain a EDC agnostic solution.
- Leverage CDISC Library where possible.

COSA - Proposal 1 - {oak} as a open-source

How ?

- Take the current Roche version of the {oak} package. Remove any Roche specific components and create an open source {oak} package
- Retain CDISC SDTM derivations, like BL Flag derivation, Visit Day, etc.
- Enhance basic algorithms to work with CDISC Library (ASSIGN_CT, HARDCODE_CT, MULTIPLE_RESPONSES, etc)
- Enhance the package to work with Clinical raw data in ODM format. (EDC agnostic)
- Template R programs to create SDTM domains.

COSA - Proposal 1 - {oak} as a open-source

Roche Involvement -

- Roche will take ownership to provide the first version of the {oak} R package (Tentative Q2 - 2023).
- It will be permissively licensed so no potential for monetization.
- We'll be advertising it via pharmaverse once open source and we'd submit it for adding to the COSA list like how they included admiral to help us raise further awareness.

The CDISC Library API is well documented. We may need support from COSA to reg CDISC libraries as needed.

COSA - Proposal 2 - PoC to automate SDTM based on CDASH standards

Vision

- Open source Metadata driven SDTM automation solution that enables data scientists to automate SDTM datasets in R.
- Enable SDTM automation when CDASH standards are adopted from CDISC Library.
- Follow ODM standards & remain a EDC agnostic solution.
- Completely leverage OAK Algorithms, CDISC Library and CDASH eCRFs.
- Provide a framework for automation when CDASH standards are extended to meet study needs.

COSA - Proposal 2 - PoC to automate SDTM based on CDASH standards

How?

- Pick simpler domains like VS, MH, CM for a PoC.
- Add algorithms and associated metadata to CDISC Library for CDASH standards. (similar to what Roche team did in Roche's MDR)
- Modify Roche version of the {oak} package to work with CDISC Library & ODM clinical data format to enable Metadata driven automation. This might be a extension of {oak} package, something like {oak.cdash}
- Use {oak.cdash} package and automate SDTM. If successful, expand to all CDASH standards

Roche Involvement - Open to collaborate with other interested parties and guide them through the PoC.

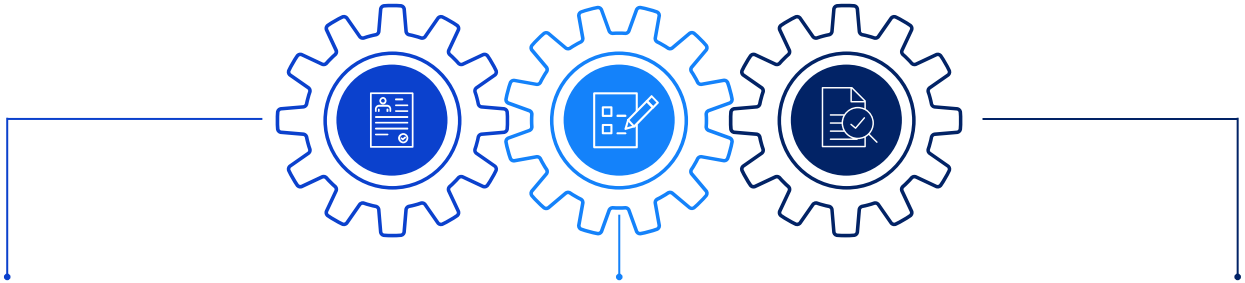
COSA - Proposal 2 - PoC to automate SDTM based on CDASH standards

COSA Support - It will be great to have COSA support to explore this PoC. We need industry support to try this option.

Roche Involvement - Open to collaborate with other interested parties and guide them through the PoC.

COSA - Proposal 3 - raw.synthetic.data

Enhance the existing Framework to



Accommodate ODM schema which can support any EDC or vendor data.

Explore ways to generate Biologically meaningful test data. Apply ML techniques and tap into publicly available data.

Define and develop guidelines for Biologically meaningful data (Ex. Lab ranges, values etc).

Currently in discussion with Pfizer, Janssen, NovoNordisk, Biogen, Teva regarding next steps.

Next steps - TBD with COSA.

COSA - Proposals - Summary

In Summary, this collaboration could enable

SDTM in R:
{oak} - Enables Data Scientists to create modularized R programs to create SDTM
{raw.synthetic.data} - Enables Data Scientist to test their R code with synthetic data

CDASH SDTM Automation PoC:
An attempt to automate SDTM creation across industry when adhering to CDASH standards.

Next Steps

Doing now what patients need next