cdisc

2023

US

INTERCHANGE

FALLS CHURCH, VA | 18-19 OCTOBER

OAK SDTM Automation: Enabling industry-level programming of SDTM in R

Presented by:
Yogesh Gupta, Sr. Director, Programming standards, Pfizer
Rammprasad Ganapathy, Principal Data Scientist, Data & Statistical Sciences, Genentech

Meet the Speakers

Yogesh Gupta

Title: Sr. Director, Statistical Data Sciences & Analytics

Organization: Pfizer

Strategic leader contributing to building large teams, plans, and roadmaps

Expertise in end-to-end standards , systems & processes

Rammprasad Ganapathy

Title: Principal Data Scientist, Data and Statistical Sciences

Organization: Genentech/Roche

Passionate about automation, with experience in statistical programming, EDC, and standards development. Enjoys R and SAS programming and leads software development projects.

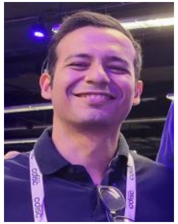
Agenda

- OAK Vision & High-Level Roadmap
- Team Structure
- Proposed R packages - {sdtm.oak}, {mint}, and {raw.synthetic.data} - Scope & Deliverables
- Onboarding Training Plan
- How to Stay Connected

Leadership Team

Charles Shadle
CDISC / COSA

Sam Hume
CDISC / COSA



Omar Garcia
CDISC

Yogesh Gupta
Pfizer

Rammprasad Ganapathy
Roche/Genentech

Bhaskar Ponugoti
Merck

Lisa Hourteloot
Pfizer

Aditya Parankusham
GSK

Susheel Arkala
Vertex

Preetesh Parikh
Pfizer

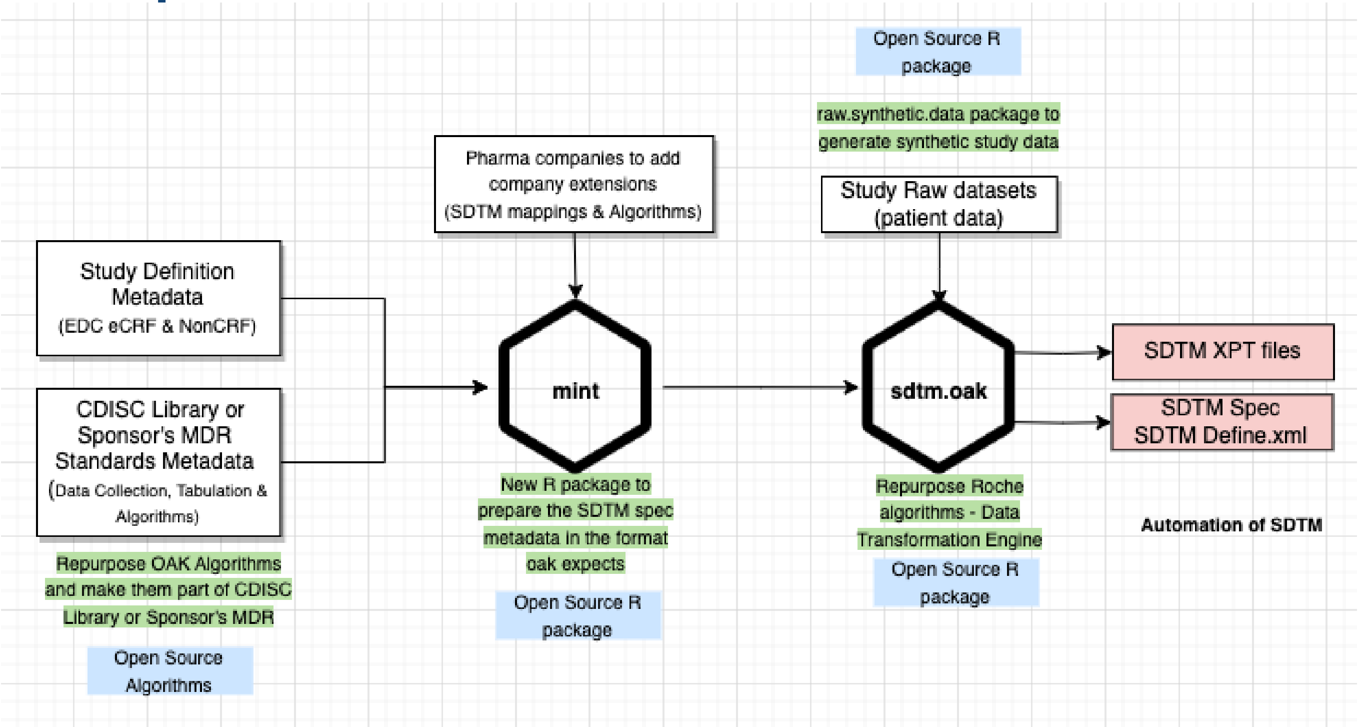
Edgar Manukyan
Roche/Genentech

Sandra VanPelt
Pfizer

Joshua Bernal
Genentech

OAK Open Source Vision

OAK Open-source SDTM Automation Vision



Open Source R package

raw.synthetic.data package to generate synthetic study data

Study Raw datasets (patient data)

Pharma companies to add company extensions (SDTM mappings & Algorithms)

Study Definition Metadata (EDC eCRF & NonCRF)

CDISC Library or Sponsor's MDR Standards Metadata (Data Collection, Tabulation & Algorithms)

SDTM XPT files

SDTM Spec SDTM Define.xml

Repurpose OAK Algorithms and make them part of CDISC Library or Sponsor's MDR

Open Source Algorithms

New R package to prepare the SDTM spec metadata in the format oak expects

Open Source R package

Repurpose Roche algorithms - Data Transformation Engine

Open Source R package

Automation of SDTM

- ❖ Metadata Driven SDTM Automation Solution
- ❖ EDC/Source and Data standards agnostic Solution (ODM Compliant – To be determined)
- ❖ Can be extended to include Sponsor extensions

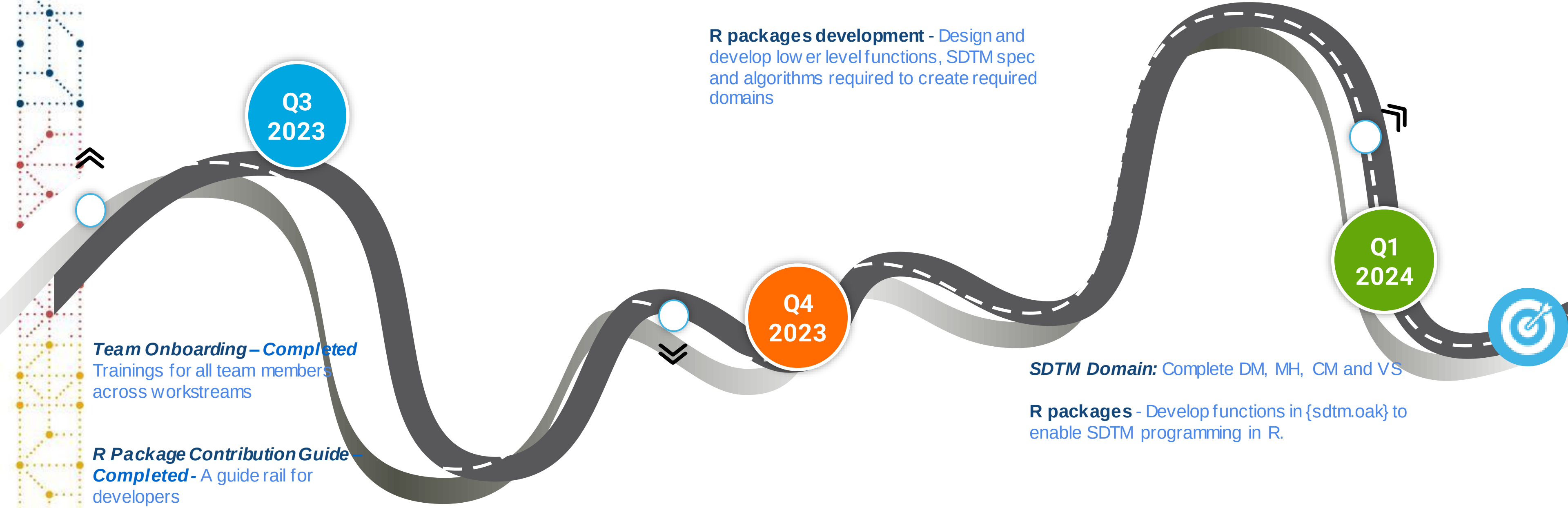
OAK and CDASH Standards

Does OAK need CDASH Standards?

- ❖ Short answer is No.
- ❖ OAK Algorithms are generic enough that can work with any data standards.
- ❖ Roche’s MDR does not have CDASH Standards. OAK is developed based on standards that are not CDASH Compliant
- ❖ For the purpose of this PoC sponsored by COSA, we are proposing to use CDASH standards.
- ❖ We are also involving the participating sponsor’s MDR and EDC systems in the development process to achieve the EDC and standards agnostic approach to automate SDTM.

COSA - OAK SDTM Automation PoC Vision

- ❖ Develop an Open-source Metadata driven R based SDTM automation solution.
- ❖ We will use CDASH aligned collection for POC, but vision is to exend this solution for any source standard (not just CDASH)
- ❖ Remain an EDC-agnostic solution.
- ❖ Completely leverage OAK Algorithms, CDISC Library.
- ❖ Provide a framework for automation when CDASH/Sponsor standards are extended to meet study specific nuances.
- ❖ Explore ways to use oak with Real World Data (RWD), external clinical trials, claims data, public consortia data



Team Onboarding – Completed Trainings for all team members across workstreams

R Package Contribution Guide – Completed - A guide rail for developers

Raw data Challenges – Decided to move forward with experimentation of long lean format for raw datasets.

SDTM Domain: Start with simple domains like DM, MH, CM and VS

Patient data in long lean format– Develop functionalities to create raw in Dataset long lean format.

R packages development - Design and develop lower level functions, SDTM spec and algorithms required to create required domains

SDTM Domain: Complete DM, MH, CM and VS

R packages - Develop functions in {sdtm.oak} to enable SDTM programming in R.

Q2-2024 and Beyond

Focus on Metadata driven automation by adding the algorithms and associated metadata to CDISC Library and develop other associated R packages.

Team Structure

OAK Team Structure

Leadership Team

- Define project scope
- Establish strategy
- Create roadmap
- Monitor progress
- Foster collaboration and effective communication

LT

Metadata Curation Team

- Develop & maintain algorithms' metadata standards
- Add algorithms metadata to MDR and CDISC Library
- Create user documentation to aid companies using approach with own metadata

MCT

OAK Development Team

- Lead effort to develop & maintain 3 OS R packages: {mint}, {oak}, & {raw.synthetic.data}
- Develop & maintain code contribution guide
- Produce unit test cases
- Prepare documentation & vignettes
- Designate sprints for further development

ODT

OAK Community Team

- Help to code / test R functions designed by the ODT
- Ensure packages meet required specifications
- Drive compatibility across different environments / platforms

OCT

Open Source OAK – Proposed R Packages

COSA – PoC Development Plan

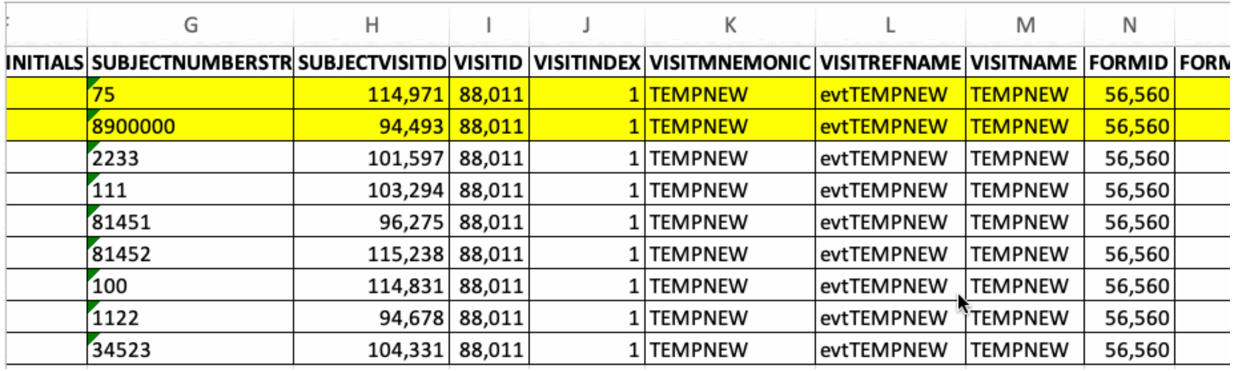
At a very high level, the open source journey can be split to three high level steps.

1. Develop R functions to pre-process Raw datasets to long lean format for easier processing.
2. Develop basic Algorithms & functions in {sdtm.oak} and enable SDTM datasets creation using R. No automation at this point. This will help growing user community and also keep them engaged.
3. Focus on Metadata driven automation by developing the R packages & metadata required for automation. Leverage CDISC Library and participating Sponsor MDR.

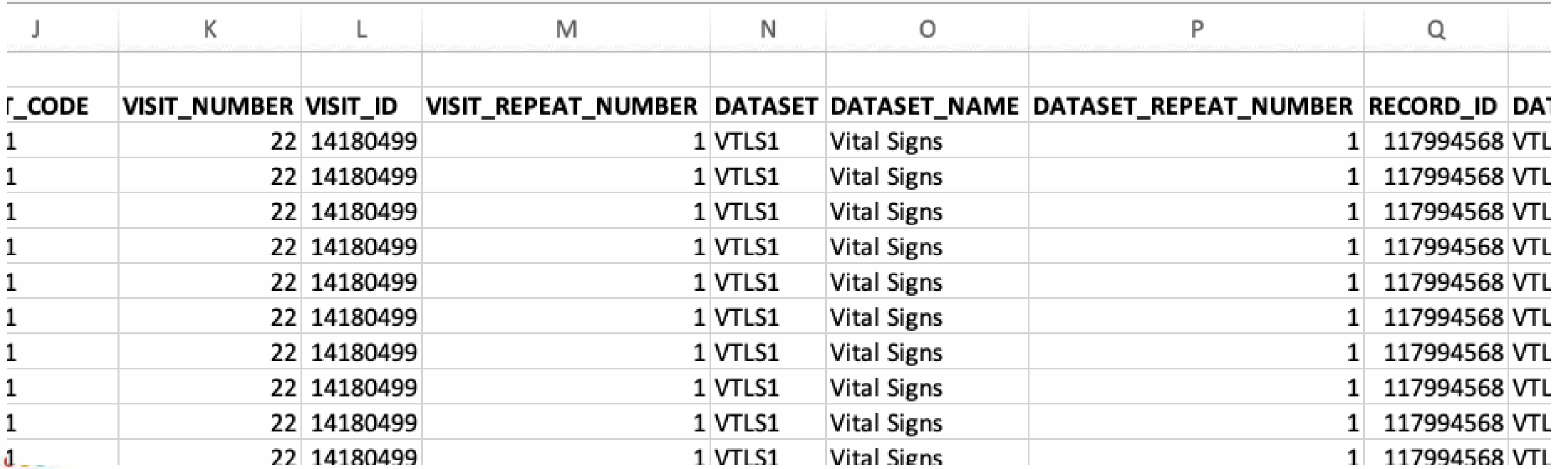
COSA - OAK – Raw data Challenges

- ❖ EDC systems are not able to provide raw data in ODM format. We had to spend quite a bit of time to figure out an alternate approach.
- ❖ EDC systems provide data in multiple formats, that it is not possible to develop a {oak} package that can be EDC agnostic.
- ❖ So we decided to pivot and first overcome the raw data issue by experimenting to pre-process any raw data from a wide format to a long-lean format.
- ❖ We will develop R functions to tackle this issue.

Wide as is Format to long_lean format (sample)



	G	H	I	J	K	L	M	N	
INITIALS	SUBJECTNUMBERSTR	SUBJECTVISITID	VISITID	VISITINDEX	VISITMNEMONIC	VISITREFNAME	VISITNAME	FORMID	FORM
	75	114,971	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	8900000	94,493	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	2233	101,597	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	111	103,294	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	81451	96,275	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	81452	115,238	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	100	114,831	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	1122	94,678	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	
	34523	104,331	88,011	1	TEMPNEW	evtTEMPNEW	TEMPNEW	56,560	



J	K	L	M	N	O	P	Q	
T_CODE	VISIT_NUMBER	VISIT_ID	VISIT_REPEAT_NUMBER	DATASET	DATASET_NAME	DATASET_REPEAT_NUMBER	RECORD_ID	DAT
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL
1	22	14180499	1	VTLS1	Vital Signs	1	117994568	VTL

{oak} – Data Transformation Engine

Scope:

- ❖ Data Transformation Engine that generates SDTM datasets. This package will have the code for the Algorithms and generic SDTM functions.
- ❖ The preliminary version of the package should enable creation of SDTM datasets in R.
- ❖ Subsequent releases should aim for automation and gets SDTM mappings metadata input and works with CDISC/Sponsor MDR Library & any patient data format to enable Metadata driven automation.

Input:

- Study SDTM mappings Metadata.
- Patient data

Output:

SDTM datasets

{mint} – Prepare SDTM mapping metadata

Scope

Prepares the study SDTM mappings metadata (i.e. specifications) for any source and data standard in the format OAK expects. Use CDISC Library to develop this New R package. This can be extended to any EDC and Sponsor MDR.

Input:

- Standards Metadata from MDR. Standards metadata will include the Algorithms associated with the SDTM mappings.
- Study metadata.

Output

Study SDTM mappings metadata in the format OAK expects.

Not in Scope:

How the data is extracted from Sponsor's MDR and Sponsor's EDC is not included. Every sponsor may need to build this extraction mechanism.

{metacore} – This feature could be implemented in the {metacore} package. Yet to the decided.

{raw.synthetic.data} – Generate raw test data

Scope:

An EDC/Source standard Agnostic solution. Some generic functions from the Roche version of the {raw.synthetic.data} package can be exported to the open-source version.

Input:

- Study metadata (eCRF, Visit definitions)

Output:

Patient data.

Not in Scope:

How the metadata is extracted from Sponsor's EDC is not included. Every sponsor may need to build this extraction mechanism.

Onboarding Training Plan

Training Plan

Self- paced Learning

- Self Paced Learning
- CDASH Standards
- CDASH | CDISC webpage

  - Specifically, the 6 short primer videos on the primer tab

- CDISC LMS (requires self registration with cdiscID)

  - CDISC for Academic Researchers (free, on-demand training)

  - CDISC for Newcomers Webinar (free, recording)

Training Plan – Pre recorded Training available for Volunteers

CDISC Training

- CDISC Library
- CDISC Biomedical concept
- Introduction to ODM
- Introduction to GitHub and Project management in GitHub

Roche Training

- Introduction to Algorithms
- Roche MDR
- SDTM Spec Metadata
- Roche {OAK} implementation

Working Methodology

Working Space: Github repository

- https://github.com/pharmaverse/sdtm.oak
- https://github.com/pharmaverse/mint
- https://github.com/pharmaverse/raw.synthetic.data
- Contribution Model similar to Admiral.
- Detailed contribution model will be shared with team
- Team members can assign issues to themselves and work on them.

HOW TO STAY CONNECTED

❖ Slack
oakgarden.slack.com

❖ Wiki
https://wiki.cdisc.org/display/oakgarden

❖ GitHub
https://github.com/pharmaverse/oak
https://github.com/pharmaverse/mint
https://github.com/pharmaverse/raw.synthetic.data

Thank You!

cdisc