Paper SI02

Oak Garden - Metadata-driven SDTM automation in R

Amy Franklin, Roche, Welwyn Garden City, United Kingdom
Magdalena Krochmal, Roche, Basel, Switzerland
Rammprasad Ganapathy, Roche, San Francisco, USA

ABSTRACT

SDTM mapping is often considered a simple process but remains a challenge across the Pharmaceutical industry. OAK is a metadata-driven SDTM automation solution that uses simple algorithms to transform source data into the target tabulation model. Built on an R-based approach, OAK automates study standards, while the same core algorithms can also be used to prepare non-standard mapping definitions based on study requirements. Its flexible architecture and modular design make it easy to program non-standards, even for complex scenarios. OAK leverages metadata and global data standards to automate approximately 80% of SDTM domains with 22 reusable algorithms. In this presentation, we will showcase Roche's implementation of OAK SDTM automation. We are also working with COSA and industry partners to make this an open-source solution, which will be covered as a separate topic.

INTRODUCTION

The Study Data Tabulation Model (SDTM) is a standard data model that defines the structure and content of clinical trial data. Established by the Clinical Data Interchange Standards Consortium (CDISC), SDTM provides a standard for organizing and formatting data to streamline data collection, analysis, and reporting processes. SDTM is one of the requirements for clinical data submission to major regulatory agencies like the Food and Drug Administration (USA) and the Pharmaceuticals and Medical Devices Agency (Japan). SDTM datasets summarize information about study subjects in the form of collections called domains. Each domain groups data by topic-specific commonality, e.g., demographics, adverse events, findings, and other relevant clinical trial data. SDTM was declared the standard format for submitting data related to clinical studies in 2004; however, preparation of the SDTM datasets remains one of the most time-consuming tasks in the clinical submission workflow.

PROBLEM STATEMENT

Automating SDTM processes can provide significant advantages to the pharmaceutical and clinical research industry, but it also presents several challenges that must be addressed. One of the challenges is ensuring data quality and consistency, which requires meticulous attention to detail to maintain compliance with SDTM standards. Additionally, complex data mapping and transformations are required to convert raw clinical and nonclinical data into SDTM-compliant datasets. Compliance with evolving regulatory standards is also crucial; automation systems must keep pace with these changes. Integrating disparate systems and data sources into a unified automation workflow can be challenging. Ensuring data accuracy and quality is essential, and automation solutions must incorporate robust validation and quality control checks. Implementing and maintaining SDTM automation systems requires significant resources and specialized skills. Converting existing legacy data into SDTM format can be time-consuming and complex, and managing metadata and updates as standards evolve is critical. Implementing changes while maintaining data integrity can also be challenging. Finally, organizations must weigh the upfront costs of implementing SDTM automation against the long-term benefits. Addressing these challenges requires careful planning, investment, and ongoing monitoring to ensure data complies with evolving standards and regulatory requirements.

BENEFITS OF AUTOMATED SDTM GENERATION

Automated SDTM generation provides several benefits to clinical data management, helping to streamline the process of transforming raw clinical trial data into the standardized SDTM format. One of the significant benefits of automated SDTM generation is the increased efficiency it brings to clinical trials. By reducing the time required to map raw clinical trial data into SDTM datasets, valuable time is saved, and the overall clinical trial process can be made more efficient. Additionally, automated data mapping ensures that the generated SDTM datasets are more accurate, eliminating the potential for human error in the data mapping process. This, in turn, ensures data consistency and integrity. Automated SDTM generation also helps to ensure compliance with regulatory standards. Since SDTM datasets are used to support regulatory submissions, it is essential to meet specific regulatory requirements. Automated SDTM generation can ensure that generated datasets conformant with regulatory standards, saving time and reducing the risk of errors. Finally, automation can result in cost savings by reducing the need for manual data management, increasing efficiency, and reducing the risk of errors. Automated SDTM generation provides several benefits for clinical trial data management and analysis, making the process more efficient, accurate, compliant, and cost-effective.

OAK GARDEN METADATA FLOW

OAK Garden is a metadata-driven SDTM automation solution. We refer to the term metadata in multiple places. “metadata” is a general concept, and “Metadata” refers to our Oak Garden-specific metadata.

In general, metadata is “data about data”. It does not include any patient-level information. The Metadata provides the blueprint of the data to be collected in a study.

The Oak Garden is made up of several tools (or plants!) that enable the flow of Metadata to enable SDTM automation (see Fig. 1)

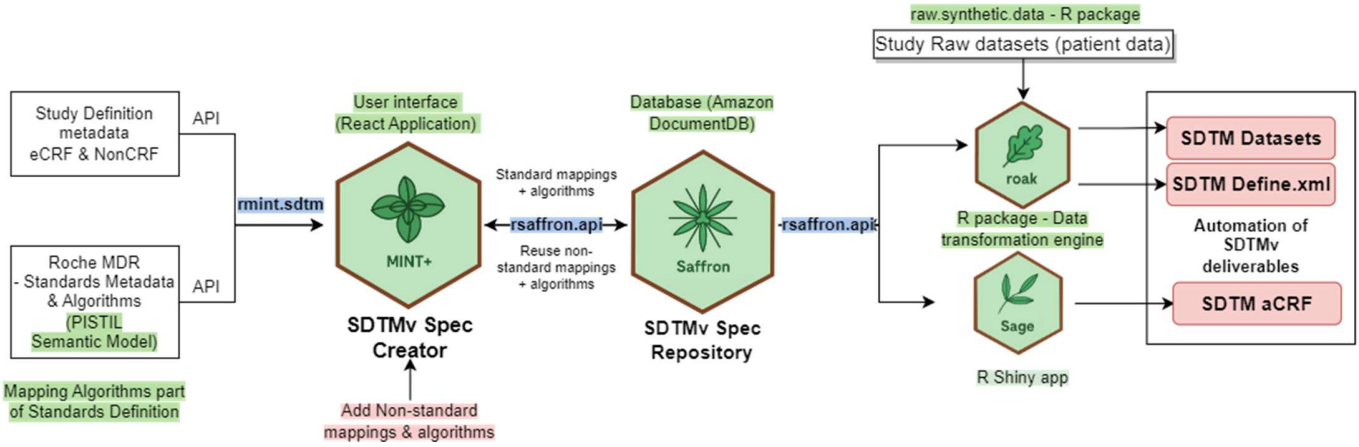

Fig.1. Metadata flow through the Oak Garden Tools

{rmint.sdtm) - METADATA CREATION

The Metadata comprises 2 parts: Study Definition Metadata and Standards Metadata.

Study Definition Metadata

Study Definition Metadata is also referred to as Study Metadata. Study Definition Metadata provides information about the eCRF (electronic Case Report Form) and NonCRF (Non-Case Report Form) data collected in the study.

eCRF Metadata

The eCRF Design Metadata is fetched from the EDC (Electronic Data Capture) system set up at the start of a study. This Metadata includes:

- Forms Metadata - Identifier (OID) and other properties of the eCRF.
- Fields Metadata - Identifier (OID), question label, datatype, and other properties of data collection fields in the study.
- Data Dictionaries - Identifier (OID) and the controlled terms collected at the source.
- Visits - Name of the visits as defined in the EDC.

NonCRF Metadata

NonCRF Metadata is the blueprint metadata that is defined by study teams with the vendor providing the NonCRF data. The blueprint metadata includes

- NonCRF dataset name.
- NonCRF variable name, datatype, and associated appendix name.
- Appendix metadata which includes the xxTEST, xxTESTCDs, and all codelist values as collected in the blueprint.

Standards Metadata

The PISTIL semantic model in Roche’s MDR (GDSR - Global Data Standards Repository) holds the Standards Metadata, which includes:

- The relationship between Data Collection Standards (eCRF & NonCRF models), SDTM mapping and Controlled Terminology

- Machine-readable standard SDTM mappings as stored in GDSR.
- Algorithms and associated Metadata that are required for the SDTM automation of standards in the study.

PISTIL is not directly used by the end users while using `roak`. The `rmint.sdtm` R package uses PISTIL and pulls the SDTM mappings, Algorithms, and the associated Metadata for the standards in the study and makes it available to users in MINT+.

As the users create the Study SDTM Mapping Specification in MINT+, `rmint.sdtm` compares the Study Definition Metadata to the Standards Metadata and prepares the SDTM Mapping Metadata required for `roak` to automate SDTM. This Metadata is stored in a machine-readable JSON format in the Saffron DocumentDB and is made available to the users in the MINT+ Data Tabulation tab. Users can also add the Non-standard mappings in MINT+, which will be stored in the Saffron DocumentDB in the same format.

Study SDTM Mappings Metadata is also called SDTM Spec or Saffron Metadata.

SDTM SPEC CREATOR - MINT+

MINT+ is a web application that enables Statistical Programmers or Data Scientists to create the Study SDTM specification. MINT+ serves as the user interface, and Saffron is the data storage. Saffron stores the Study SDTM specification created by Statistical Programmers or Data Scientists in a machine-readable JSON format. Every user action in MINT+ is stored in the Saffron database.

Please read *Paper AD13 - MINT+ Automation and Digitization of SDTM Specification* for more information on MINT+.

SAFFRON - SDTM SPEC REPOSITORY

Saffron is a database for MINT+ where every user action is stored. The endpoints in Saffron enable the data scientists to store and read the SDTM mapping specifications or metadata in a machine-readable format.

{roak} - DATA TRANSFORMATION ENGINE

`roak` is designed based on reusable algorithms that aim to automate the study's SDTM deliverables in R. `roak` is the data transformation engine and is a key part of the OAK Garden - SDTM automation solution from Project Botany.

How it works?

`roak` is created as an R package, following the below concept:

- One mapping function for each mapping algorithm.
- Functions to perform generic tasks like converting to ISO date, derive study day, etc.
- `roak` is programmed in a generic way so that the functions can be used in any study.
- `roak` R package requires other components of the OAK Garden to automate SDTM mapping in a study.

Input required for `roak`

- SDTM mapping metadata from Saffron includes the SDTM mapping for standards in the study with the algorithms from GDSR and user-entered SDTM mapping for Non-standards. It also comprises algorithms for non-standards as created by data scientists in MINT+.
- Study configuration files created by a user that provide information about the study and are required for data transformation.
- Study raw datasets are eCRF and non-CRF raw datasets.

Output generated by `roak`

Study SDTM eSubmission Deliverables

- SDTM datasets
- SDTM controlled terminology
- SDTM value-level metadata
- SDTM define.xml

In summary, all SDTM eSUB deliverables are part of the SDTM automation solution.

REUSABLE ALGORITHMS

CORE CONCEPT

SDTM mappings are defined as algorithms that transform the collected (eCRF, nonCRF) source data into the target SDTM data model. Mapping algorithms are the backbone of the OAK Garden - SDTM automation.

22 unique mapping algorithms accommodate mapping of all the TA (Therapeutic Area) eCRF standards & Non-CRF data models. Data collection standards in the GDSR (PISTIL semantic model) have over 13500 unique SDTM mappings with pre-defined algorithms. The algorithms and associated metadata drive the automation of SDTM.

Key Points:

- Algorithms can be re-used across multiple SDTM domains.
- Algorithms are pre-specified for data collection standards in GDSR.
- Users can reuse/add the same algorithms for non-standards in the study. Both standards and non-standards can be supported.
- Programming language agnostic - this concept does not rely on a specific programming language for implementation. The OAK team implemented them as R functions.

Here is an example of reusing an algorithm across multiple domains, variables, and also to a non-standard:

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

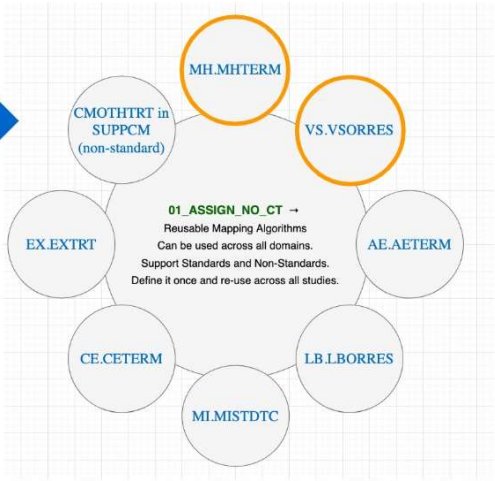

Fig.2. Example of reusable algorithms and application in SDTM mappings.

SUB-ALGORITHMS

OAK Garden supports two levels for defining algorithms. For example, there are some SDTM mappings where a certain action has to be taken only when a condition is met. In such cases, the primary algorithm checks for the condition, and the sub-algorithm executes the mappings when the condition is met.

Currently, sub-algorithms must be provided for these main algorithms.

- 09_IF_THEN_ELSE
- 12_MAP_TOPIC
- 07_DATASET_LEVEL

Algorithms can be interchangeably used as algorithms and as sub-algorithms, as seen in Fig.3 (not an exhaustive list):

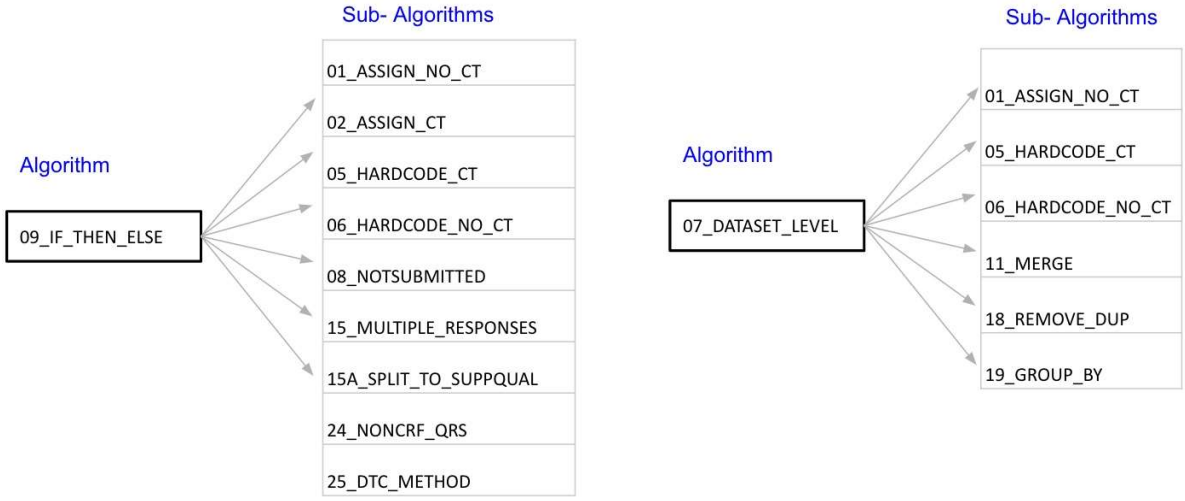

Fig.3. Example of sub-algorithms and how they connect to core algorithms

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

NON-STANDARD MAPPINGS

A study can have standard eCRF and non-CRF models that are sourced from GDSR and also study specific eCRF, and non-CRF elements that are required as per the study protocol. OAK Garden automates the study standards and aims to provide tools necessary to program non-standards in the study, as in Fig.4.

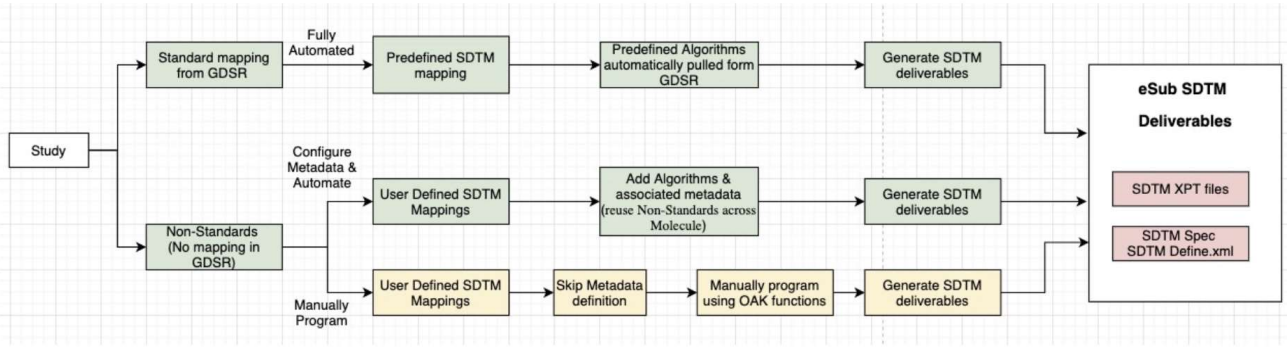

Fig.4. The 3 routes to performing SDTM mappings

STANDARDS IN THE STUDY

The algorithms and the associated metadata required for data transformation are added to the Pistil semantic model in GDSR. OAK Garden can automate the standard SDTM mappings in the study out of the box.

HANDLING NON-STANDARDS

Non-standards in the study can be handled in two ways, as listed below.

Configure Metadata and Automate

OAK algorithms can be used to automate non-standards using the intuitive point-and-click user interface in MINT+.

Program Non-Standards

The non-standards in the study can be programmed manually. An SDTM domain is first created by `roak,` which includes the standards in the study, and then the non-standards can be added to the domain using basic R programming.

STUDY TEAMS' FEEDBACK SUMMARY

All of the study teams foresee OAK automating the SDTM setup at the study level. The feedback from the study teams is below.

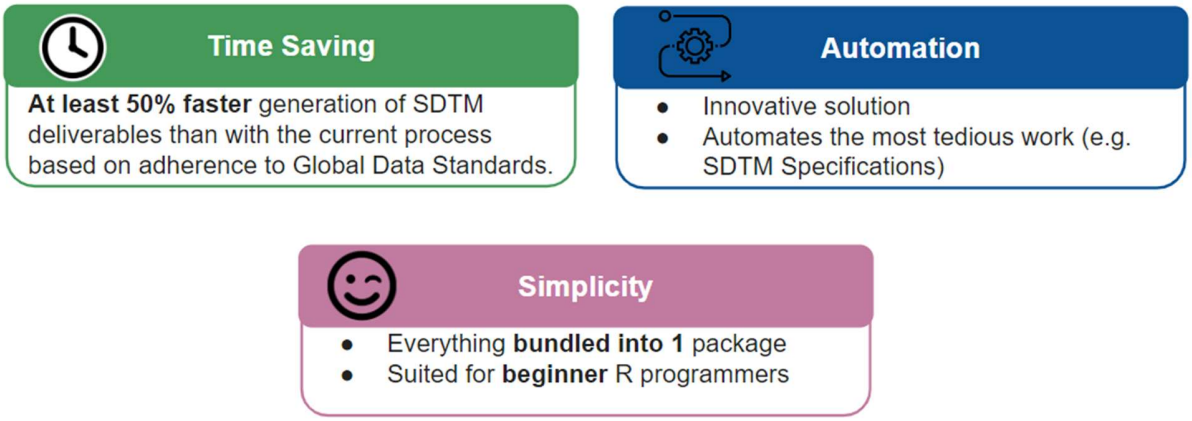

Fig.5 Feedback from Early Adopters about the Oak Garden

OPEN-SOURCING OAK

We have kicked off the OAK (R-based, metadata-driven, EDC agnostic) Open Source SDTM automation COSA-sponsored initiative and the response has been incredible. We have 70+ Volunteers from the Pharmaceutical industry who have joined us in this exciting endeavor. Stay tuned for updates as we continue to achieve milestones and reach new heights. Please find below the URLs for the description of OAK, as well as the slides and video recording of the kick-off.

Description of OAK: https://lnkd.in/gsQbQkKx

Slides from the kick-off: https://lnkd.in/gfa8g9Wq

Video recording of the kick-off: https://lnkd.in/gvGMeqin

Fig.6 shows the open-source OAK Vision, you will see it is very similar to the Roche Oak Garden, just more generic and simpler to work with different companies' data.

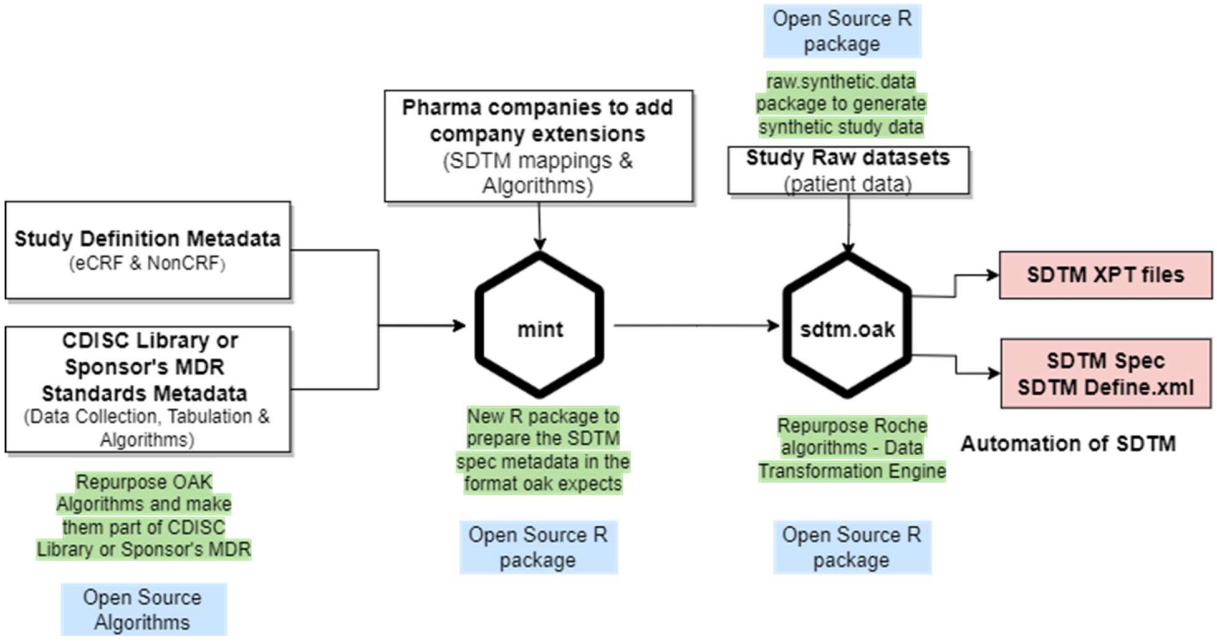

Fig.6 Open Source Oak Vision

Thanks to the leadership team for driving this! See Fig.7 for the broad depth of knowledge and experience leading this initiative.

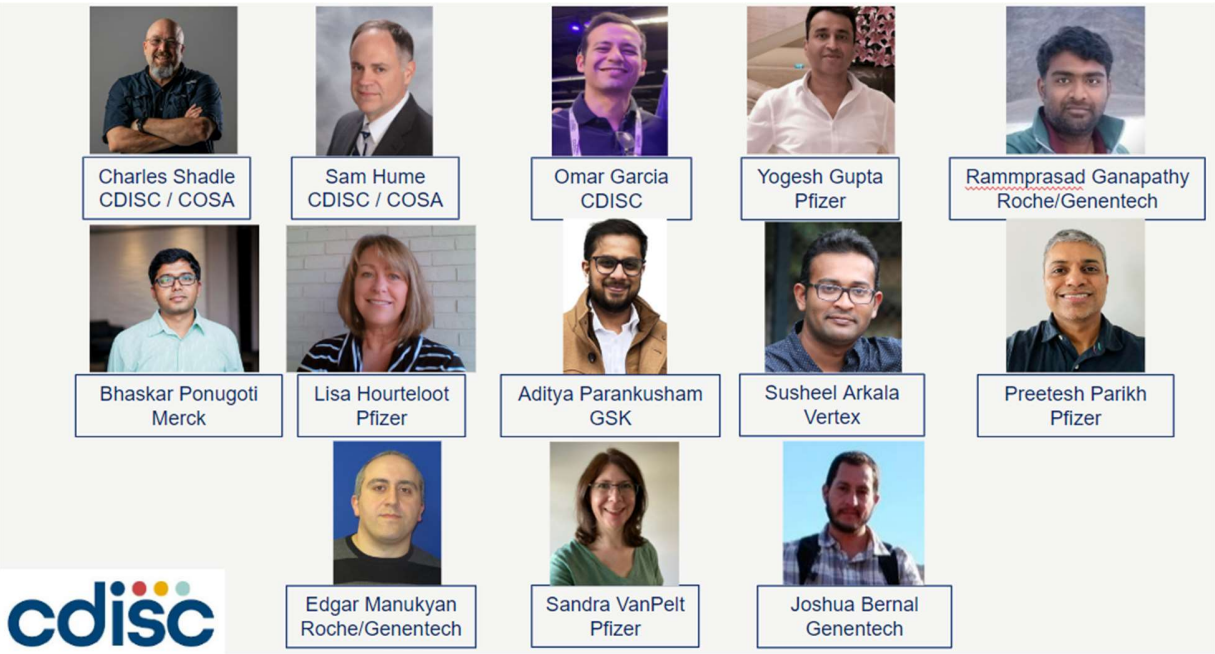

Fig.7 Open Source Oak Leadership Team

CONCLUSION

SDTM automation is becoming increasingly popular in the pharmaceutical industry as it helps to streamline the process of creating SDTM datasets. This can help reduce costs, save time, and improve compliance with regulatory requirements. The Oak Garden is showing this is possible and improving how we create our SDTM deliverables in Roche. The metadata-driven approach utilizing reusable algorithms enables the model to be flexible and programming language agnostic. Standard and Non-Standard mappings are able to be automated with significantly less hands-on programming needing to be performed by study teams, saving time and money. Many study teams have started to use the Oak Garden within Roche, and the feedback gathered from our Early Adopters speaks for itself as to how influential this tool is. So much so that we are embarking on a mission to create an Open-Source version of the Oak Garden to revolutionize how we create SDTMs in Roche and beyond, watch this space.

ACKNOWLEDGMENTS

The work presented in this paper is the result of the collaborative efforts of the authors together with the entire Oak Garden Team.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the authors at:
amy.franklin@roche.com, magdalena.krochmal@roche.com, ganapathy.rammprasad@gene.com