

Getting Started with the New CDISC Analysis Results Standard

Bess LeRoy, CDISC, Austin, USA

Richard Marshall, CDISC

Bhavin Busa, Clymb Clinical

ABSTRACT

The CDISC Analysis Results Standard (ARS) Team has developed standards to improve and facilitate the automation, reproducibility, reusability, and traceability of analysis results (Tables, Figures and Listings, aka TFL). The team created a logical model to incorporate the elements for both analysis results and associated metadata. This paper provides an overview of the CDISC Analysis Results logical model, including its background and development.

INTRODUCTION

Analysis results play a crucial role in the drug development process, providing essential information for regulatory submission and decision-making. However, the current state of analysis results reporting is suboptimal, with limited standardization, lack of automation, and poor traceability. Currently, analysis results (tables, figures, and listings) are often presented in static, PDF-based reports that are difficult to navigate and vary between sponsors. Moreover, these reports are expensive to generate and offer limited reusability. To address these issues, the CDISC ARS team has developed a standard to support consistency, traceability, and reuse of results data.

BACKGROUND

Currently, analysis results are created as static PDF documents that may contain hundreds of tables. These tables are often difficult to navigate and there is significant variability between sponsors. Generating these reports is expensive and they are typically only used once, offering limited reusability.

The current workflow of generating analysis results involves the end user generating the Analysis Data Model (ADaM) dataset, followed by generating the display in a static format such as RTF or PDF using the ADaM dataset. The Analysis Results Metadata (ARM) for define.xml is retrospectively generated, which provides high-level metadata about analysis displays and results; however, there has been no formal model or structures to describe analysis results metadata and analysis results data which leaves a gap in standardization. Our vision for the future state of analysis results reporting is a world where analysis results are machine-readable, easily navigable, and highly reusable. To achieve these goals, the ARS team has been working toward developing a logical model to fully describe analysis results metadata. This logical model will enable the implementation of a technical specification and an analysis results data framework to support reuse and reproducibility of results data.

PROPOSED APPROACH

To address the current limitations of analysis results and associated metadata reporting, we are proposing a new workflow that shifts the focus from retrospective reporting to prospective planning. Specifically, we propose that end-users generate an analysis results metadata technical specification prior to generating a display, rather than after the display has been created. This approach will allow for better planning and standardization of the analysis process, resulting in more consistent and traceable reporting.

The proposed workflow involves several steps. First, the end-user will develop the technical specification, which will include metadata about the statistical methods, data sources, and displays to be generated. Once the technical specification has been developed, the end-user will use it to generate an analysis dataset, which will contain the results data needed to generate the display. The analysis dataset will be

designed to support reuse and reproducibility of the results data, enabling more efficient and effective analysis reporting.

Finally, the machine-readable analysis dataset serves as the 'single source of truth' capturing the analysis results metadata and results data in a standardized format. This analysis results dataset can then be used to generate displays from multiple reporting purposes, such as traditional analysis reporting for the clinical study report (CSR), in-text tables for the CSR, safety reporting, meta-analyses, dynamic applications, ClinicalTrials.gov, publications, and presentations. This streamlined approach ensures consistency and accuracy in the generation of displays across various deliverables, making it more efficient and reliable for reporting and communication of analysis results.

Overall, this new workflow will enable end-users to generate analysis results metadata prospectively, with greater standardization, consistency, and traceability of analysis results reporting, enabling better decision-making and regulatory submissions. By shifting the focus from retrospective reporting to prospective planning, we believe that this approach will help to address many of the current limitations of analysis results reporting and support the development of more efficient and effective analysis standards.

In our vision for the future state, we anticipate the availability of open-source or community tools, such as the TFL Designer Community ^[1], in the industry. These tools will empower users to create machine-readable analysis metadata, which can automate the generation of analysis results data and displays. We also hope that such a tool can seamlessly integrate with existing analysis programs and report creation tools, enabling an end-to-end automation of the analysis and reporting process.

LINKML FOR LOGICAL MODEL

Many of the same metadata components are needed both to create a prospective technical specification of analyses to be performed and to give context to reported results. The CDISC ARS team developed a single logical model comprising components needed to specify analyses, to represent contextualized results and to indicate how the results are displayed. The logical model is being developed using LinkML. LinkML is an open-source schema development language and framework for generating machine-readable models ^[2]. The LinkML Generator framework generates downstream artifacts, including JSON-Schema, ShEx, RDF, OWL, GraphQL, and SQL DDL.

The logical model acts as a blueprint or a set of rules that describes how things should be organized and structured. In our case, it is being used to create a model that describes how analysis results data should be organized and structured. LinkML has allowed the creation of a standardized and consistent approach to describing analysis results data, enabling greater interoperability and collaboration across different stakeholders. Additionally, it has the added benefit of being able to convert them into various machine-readable formats such as JSON, YAML, OWL, or XML.

Furthermore, LinkML is a flexible and extensible tool, meaning that the model can be easily modified as needed to incorporate new data elements or requirements. By creating a machine-readable model, analysis results data can be more easily understood, shared, and reused by both humans and machines. LinkML will also support the development of validation rules that can be used to ensure the integrity and quality of the data, which is essential in our highly regulated pharmaceutical industry.

ANALYSIS RESULTS STANDARD MODEL

An Analysis Results Standard model refers to the framework or structure used to store and organize the information related to an analysis result. This includes both the results data generated by the analysis and the metadata that describes the characteristics of that results data.

The metadata in this model includes information about the source of the results data, the method used for analysis, and any relevant parameters or queries that were used. This metadata is important for helping others understand and reproduce the analysis result, and it can also be used for quality control purposes.

The model defines a set of classes and their attributes, which can be used to represent different aspects of the ARM-TS and ARD. The model includes classes such as "ReportingEvent", "Analysis", "AnalysisSet", "AnalysisGroup", "Output", and "GroupingFactor", among others. Each class has a set of attributes that define its properties, such as "id", "label", "condition", "whereClauses", "methods", "fileSpecifications", and so on.

The "ReportingEvent" class serves as the root of the model and includes many other classes and attributes that help to organize the data into different categories and groupings. Each reporting event represents a set of analyses and outputs created to meet a specific reporting requirement, such as a CSR or interim analysis. As shown in Figure 3, the ReportingEvent class has a set of attributes that can be used to define, itemize, and categorize the analyses and outputs needed for the reporting event:

analysisSets:

The analysis sets defined for the reporting event. Each analysis set, or subject population, is defined using the "AnalysisSet" class as a set of subjects whose data are to be included in the main analyses. This is as defined in the statistical section of the protocol.

dataSubsets:

Any defined data subsets that are used to restrict the records that are included in any given analysis (e.g., inclusion of only treatment-emergent adverse events in an AE summary analysis). Each data subset is defined using the "DataSubset" class.

analysisGroupings:

Characteristics used to subdivide the subject population (e.g., treatment, sex, age group). Each analysis grouping is defined using the "SubjectGroupingFactor" class and may either contain a set of specified groups (e.g., "Male" and "Female" for a sex grouping) or be defined as "data- driven", where the group values are obtained from the data (e.g., system organ class or preferred term for adverse events).

methods:

Defined methods used to analyze any analysis variable. Each method, which is defined using the AnalysisMethod class, may contain multiple individual operations. For example, a "summary of categorical variable" method might include separate operations for "count of subject by group" and "percent of subjects by group".

analyses:

The analyses defined for the reporting event. Each analysis is described, using the Analysis class, as a set of operations (i.e., a referenced "AnalysisMethod") performed on a specified analysis variable for any specified subset of data records (i.e., any referenced "DataSubset") for a given subject population (i.e., a referenced "AnalysisSet") which may be subdivided by one or more factors (i.e., any referenced "SubjectGroupingFactor" or "DataGroupingFactor"). The results of each analysis are contained within the "Analysis" class as a property of the analysis and each result is associated with references to all analysis components needed to give the result context (i.e., the operation used to calculate the result and a specific value for each grouping factors used for the analysis such as treatment, sex, age group, etc.)

outputs:

The outputs defined for the reporting event. Each output is defined using the Output class and may contain one or more defined displays.

listOfPlannedAnalyses:

A structured list of the analyses defined for the reporting event. The list is defined using the NestedList class, which allows referenced analyses (and outputs) to be organized into sections and sub-sections.

listOfPlannedOutputs:

An optional separate structured list of the outputs defined for the reporting event. The list is defined using the NestedList class, which allows referenced outputs (and analyses) to be organized into sections and sub-sections.

analysisCategorizations:

Any categorizations used to tag analyses or outputs. For example, a “Type of analysis” categorization could be created to tag defined analyses or outputs as “Subject Population”, “Safety” or “Efficacy”, and/or an “Estimand Analysis Type” categorization could be created to tag analyses as “Primary Estimator”, “Sensitivity Analysis”, “Supplementary Analysis”. Each categorization is defined using the “AnalysisCategorization” class.

For a more information on the ARS model, please visit <https://www.cdisc.org/events/webinars/public> where you can view previous presentations that provide an in-depth overview of the model.

CONCLUSION

The ARS team has been actively working on creating new analysis results standards. We have discussed the current state, future state, and logical model for these standards in this paper. We are excited to announce that the first version of the standards will be released in the spring of 2024. Next steps include expanding informative content through additional examples.

REFERENCES

1. (TFL Designer GitHub, 2023): <https://github.com/bhavinbusa/tfldesigner>
2. (CDISC Analysis Results Standard GitHub, 2023): <https://github.com/cdisc-org/analysis-results-standard>

ACKNOWLEDGMENTS

The author would like to extend our heartfelt thanks to the entire CDISC Analysis Results Standard team, whose hard work, dedication, and expertise have been instrumental in developing this new analysis results standard. Your contributions have been invaluable in helping to improve the standards and underlying model, and we are grateful for your collaboration and commitment to advancing the field.

CONTACT INFORMATION

Bess LeRoy

ARS Co-Lead
Head of Standards Innovation, CDISC
bleroy@cdisc.org

Bhavin Busa

ARS Product Owner & Co-Lead, CDISC Principal & Co-founder, Clymb Clinical
bhavin@clymbclinical.com

Richard Marshall

Principal Data Modeler
rmarshall@accuratesystems.co.uk