Automating ADaM Specifications Using CDISC ARS Metadata & Gen-AI

June 11th, 2025 – PHUSE SDE Boston 2025

Anna Yaggi, TFL Designer Technical SME and Statistical Programmer

Shivani Gupta, Director of Statistical Programming

CLYMB CLINICAL

Outline

- Introduction to ARS Metadata
- Proof of Concept
- Learnings
- Key Takeaways + Next Steps

CLYMB CLINICAL

CDISC Analysis Results Standard - a Logical Data Model

cdisc
Clear data. Clear impact

Analysis Results Standard (ARS) v1.0

cdisc ARS ANALYSIS RESULTS STANDARD

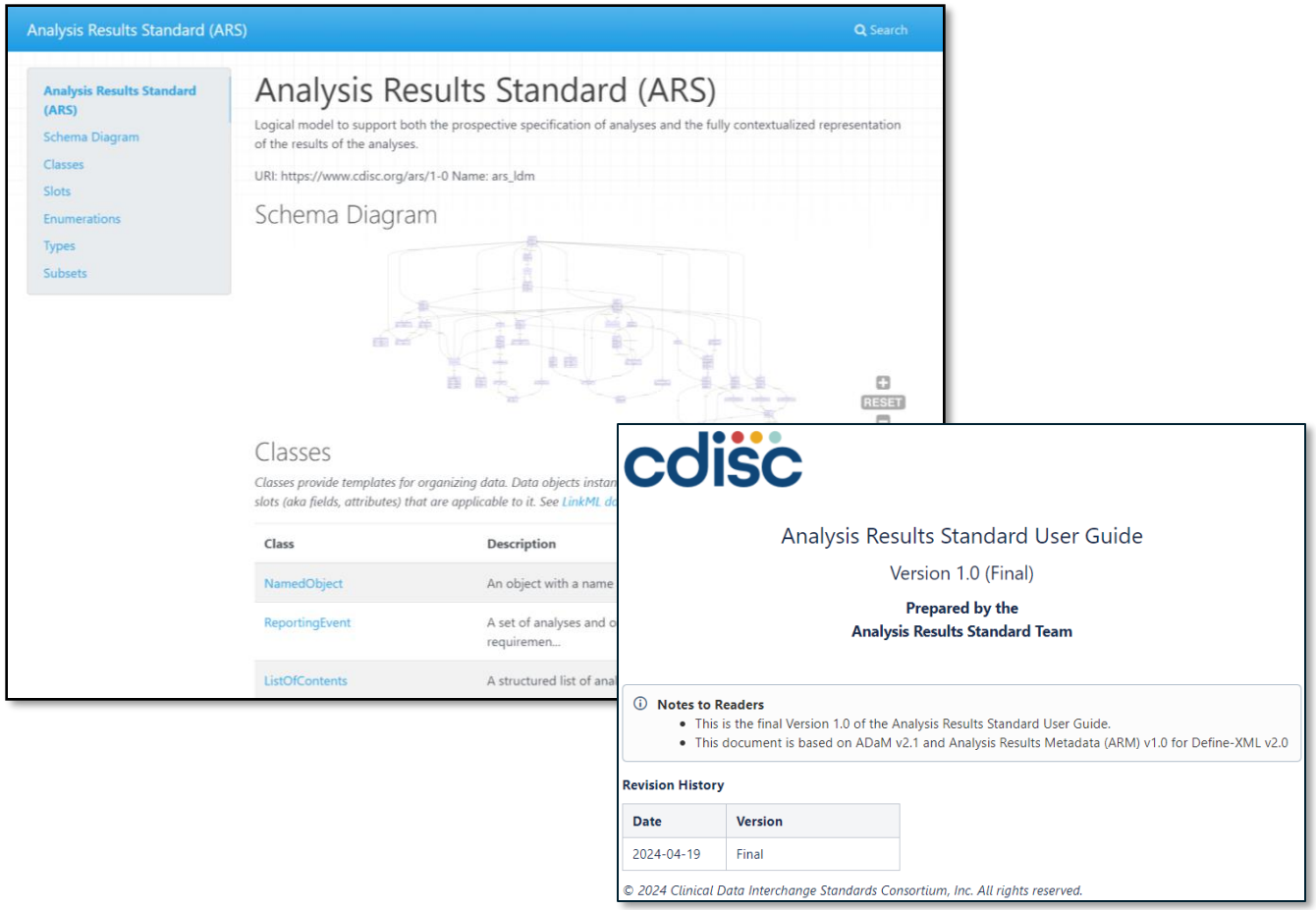
Large trials generate many analysis results in the form of tables, figures, and written reports, yet these results are rarely output in a form that is machine-readable. Previously, there has been no standard way of describing and organizing these results, making it difficult to automate their generation, make them reproducible, trace their origin, or enable them to be reused in other outputs.

To address these inefficiencies, CDISC has developed the Analysis Results Standard (ARS), which aim to facilitate automation, reproducibility, reusability, and traceability of analysis results data.

Features of ARS v1.0

- A Logical Data Model that describes analysis results and associated metadata.
- A User Guide to illustrate and exercise the model with common safety displays.

https://cdisc-org.github.io/analysis-results-standard/



https://wiki.cdisc.org/display/ARSP/Analysis+Results+Standard+User+Guide+v1.0

CLYMB CLINICAL

Model Components

Analysis Set | Data Subset | Analysis Grouping | Analysis Method | Analysis | Results

Summary of Demographics

Study – CDISC 360 Page x of y

Table 14.1.1
Summary of Demographics
Safety Population

Characteristics	Placebo (N=XX)	Xanomeline Low Dose (N=XX)	Xanomeline High Dose (N=XX)
Age (years)			
n	XX	XX	XX
Mean (SD)	XX.X (XX.XX)	XX.X (XX.XX)	XX.X (XX.XX)
Median	XX.X	XX.X	XX.X
Q1, Q3	XX.X, XX.X	XX.X, XX.X	XX.X, XX.X
Min, Max	XX, XX	XX, XX	XX, XX
Age Group, n (%)			
< 65 years	XX (XX.X)	XX (XX.X)	XX (XX.X)
≥ 65 years	XX (XX.X)	XX (XX.X)	XX (XX.X)
Gender, n (%)			
Male	XX (XX.X)	XX (XX.X)	XX (XX.X)
Female	XX (XX.X)	XX (XX.X)	XX (XX.X)
Ethnicity, n (%)			
Hispanic or Latino	XX (XX.X)	XX (XX.X)	XX (XX.X)
Not Hispanic or Latino	XX (XX.X)	XX (XX.X)	XX (XX.X)

Output

Source dataset: adsl, Generated on: DDMONYYYY:HH:MM
Program: <pid>.sas, Output: <pid><oid>.rtf, Generated on: DDMONYYYY:HH:MM

Summary of TEAE by SOC and PT

Study – CDISC 360 Page x of y

Table 14.3.1.1
Summary of TEAE by System Organ Class and Preferred Term
Safety Population

System Organ Class Preferred Term [a], n (%)	Placebo (N=XX)	Xanomeline Low Dose (N=XX)	Xanomeline High Dose (N=XX)
Number of subjects with at least one event	XX (XX.X)	XX (XX.X)	XX (XX.X)
<SOC 1>	XX (XX.X)	XX (XX.X)	XX (XX.X)
<Preferred Term 1>	XX (XX.X)	XX (XX.X)	XX (XX.X)
...	XX (XX.X)	XX (XX.X)	XX (XX.X)
<Preferred Term n>	XX (XX.X)	XX (XX.X)	XX (XX.X)
<SOC 2>	XX (XX.X)	XX (XX.X)	XX (XX.X)
<Preferred Term 1>	XX (XX.X)	XX (XX.X)	XX (XX.X)
...	XX (XX.X)	XX (XX.X)	XX (XX.X)
<Preferred Term n>	XX (XX.X)	XX (XX.X)	XX (XX.X)

Notes: TEAE=Treatment-Emergent Adverse Events.
Subjects are counted once within each system organ class and preferred term.
[a] All investigators adverse events were coded using MedDRA version xx.x.

Source dataset: adae, Generated on: DDMONYYYY:HH:MM
Program: <pid>.sas, Output: <pid><oid>.rtf, Generated on: DDMONYYYY:HH:MM

Accessing CDISC ARS

- Machine-Readable JSON Format
  - CDISC ARS metadata is provided in a structured, machine-readable JSON format.
  - This allows for efficient integration and downstream processing.

- Easily Visualized in Excel

  - The JSON can be converted into an Excel workbook for human-friendly inspection.
  - Each ARS class appears as a separate tab in the spreadsheet.
- Download via TFL Designer (Community Version)
  - You can build shells and download their full ARS metadata using the community edition of TFL Designer

CDISC ARS Metadata Model Components

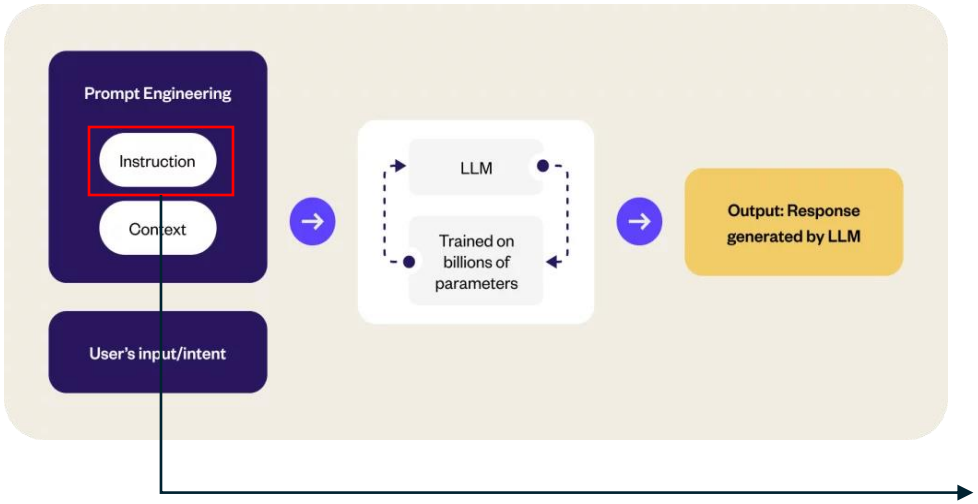
- Example information: Analysis Groupings

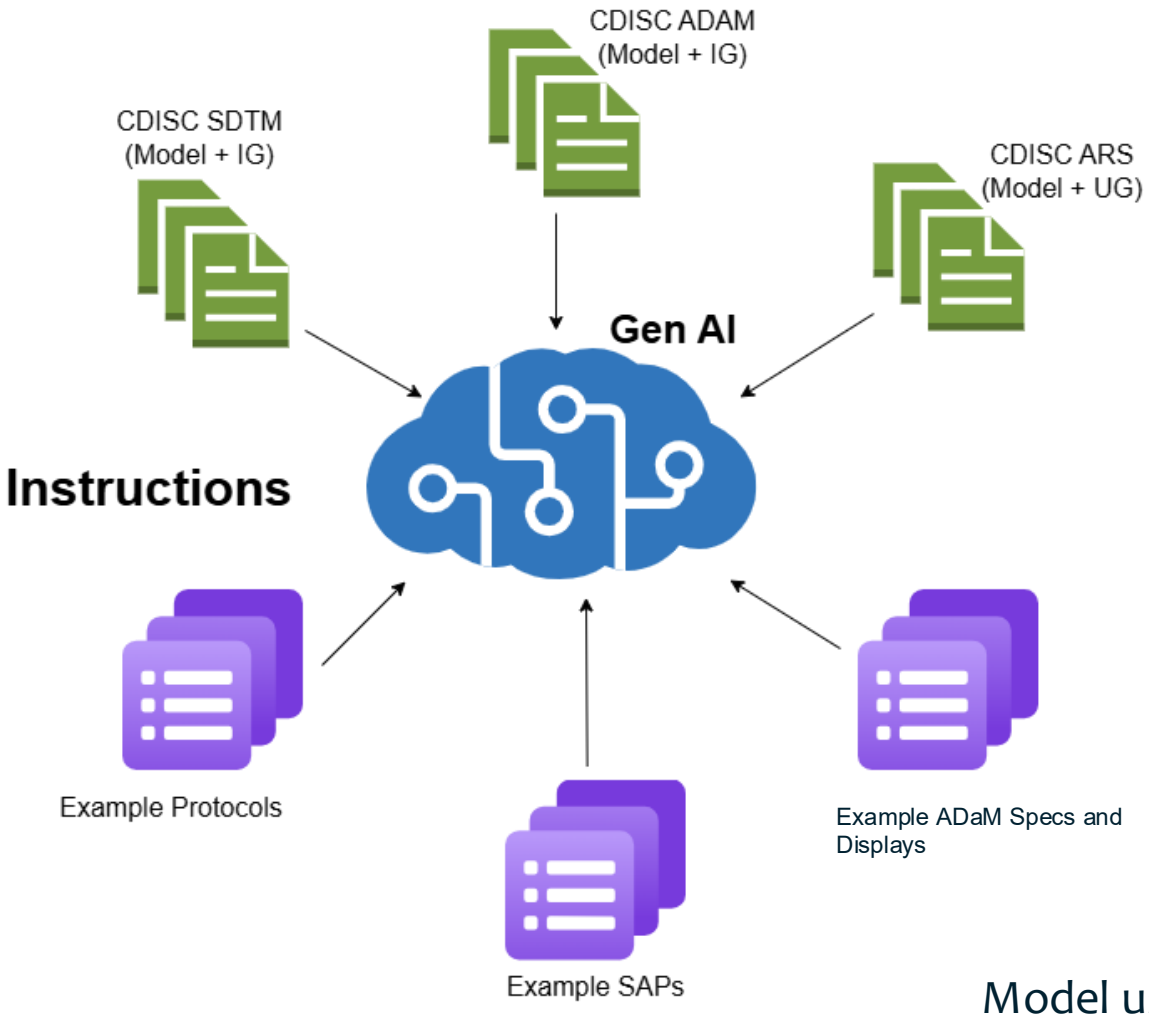
group_condition_dataset	group_condition_variable	group_condition_comparator	group_condition_value
ADSL	TRT01A	EQ	Placebo
ADSL	TRT01A	EQ	Xanomeline Low Dose
ADSL	TRT01A	EQ	Xanomeline High Dose
ADSL	SEX	EQ	M
ADSL	SEX	EQ	F
ADSL	AGEGR1	EQ	<65
ADSL	AGEGR1	IN	65-80 | >80
ADSL	RACE	EQ	AMERICAN INDIAN OR ALASKA NATIVE
ADSL	RACE	EQ	ASIAN
ADSL	RACE	EQ	BLACK OR AFRICAN AMERICAN
ADSL	RACE	EQ	NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER
ADSL	RACE	EQ	WHITE
ADSL	RACE	EQ	MULTIPLE
ADSL	RACE	EQ	NOT REPORTED
ADSL	RACE	EQ	UNKNOWN
ADSL	RACE	EQ	OTHER
ADSL	ETHNIC	EQ	HISPANIC OR LATINO
ADSL	ETHNIC	EQ	NOT HISPANIC OR LATINO

Proof of Concept

Using Generative AI to take in ARS metadata and deliver ADaM Specifications.

Process Flow for Use Case



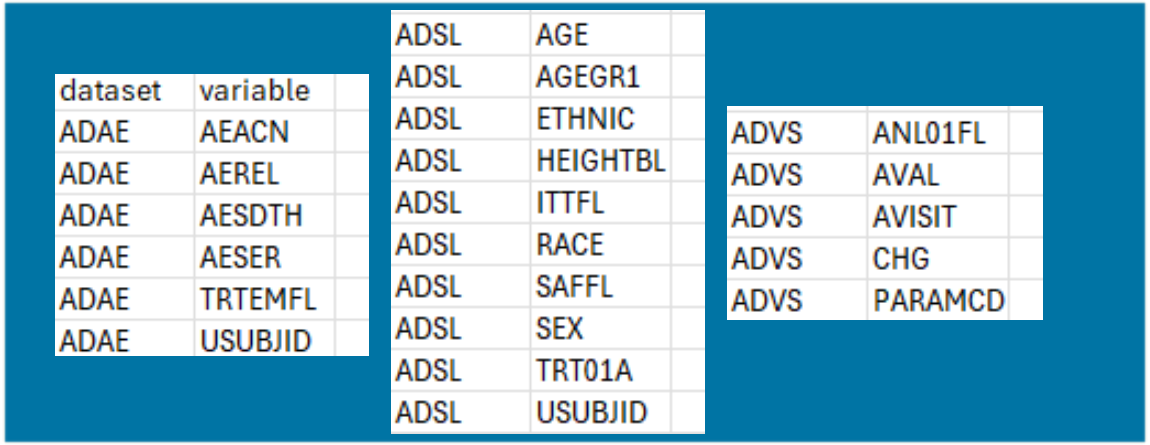


Model used GPT-4o

Use Case: Step 1

- **Prompt:**
  - I am uploading some CDISC ARS metadata (Safety_Displays_ARS) as well as the table shells associated with the metadata (Safety_Displays_Shells). First, I want you to understand this metadata.
  - Please use information from the Data Subsets tab, Analysis Sets tab, Analysis Groupings tab, and Analyses tab to come up with a list of ADaM datasets and the variables used.

- **Results:**



dataset	variable
ADAE	AEACN
ADAE	AEREL
ADAE	AESDTH
ADAE	AESER
ADAE	TRTEMFL
ADAE	USUBJID

ADSL	AGE
ADSL	AGEGR1
ADSL	ETHNIC
ADSL	HEIGHTBL
ADSL	ITTFL
ADSL	RACE
ADSL	SAFFL
ADSL	SEX
ADSL	TRT01A
ADSL	USUBJID

ADVS	ANL01FL
ADVS	AVAL
ADVS	AVISIT
ADVS	CHG
ADVS	PARAMCD

  - It seems that giving it a step to understand the metadata, as well as providing it with associated shells helps to improve accuracy.

Use Case: Step 2

Prompt:

- Thank you! Now, please refer to the **ADaM Implementation Guide** I am uploading, and train yourself on the required and common variables for these datasets. Please provide a list of required variables for the ADaM datasets you specified in the previous section.

Results:

1. ADSL (Subject-Level Analysis Dataset)

As per ADaMIG Section 3.2, required variables include:

Variable	Label
`STUDYID`	Study Identifier
`USUBJID`	Unique Subject Identifier
`SUBJID`	Subject Identifier for the Study
`SITEID`	Study Site Identifier
`AGE`	Age
`AGEU`	Age Units
`SEX`	Sex
`RACE`	Race
`ARM`	Description of Planned Arm
`TRT01P`	Planned Treatment for Period 01
≥1 Population Flag	e.g., `SAFFL`, `ITTFL`, `FASFL` (depends on SAP)

2. BDS (Basic Data Structure – applies to ADAE, ADVS, etc.)

As per ADaMIG Section 3.3:

Variable	Label
`STUDYID`	Study Identifier
`USUBJID`	Unique Subject Identifier
`PARAM`	Parameter
`AVAL` / `AVALC`	Analysis Value (numeric or character)
`ANLzzFL`	Analysis Flag (at least one recommended)
`TRTP` / `TRTA`	Planned / Actual Treatment for the record
`ADT`	Analysis Date
`ADY`	Analysis Relative Day
`ASEQ`	Analysis Sequence Number

Use Case: Step 3

- **Prompt:**
  - Please combine these two lists – the list of variables from the metadata, and the list of required variables- to create a comprehensive list of the variables that would be in these datasets.

- **Results:**

dataset	variable	source
ADAE	ADT	ADaMIG Required
ADAE	ADY	ADaMIG Required
ADAE	AEACN	Metadata
ADAE	AEREL	Metadata
ADAE	AESDTH	Metadata
ADAE	AESER	Metadata
ADAE	ANLzzFL	ADaMIG Required
ADAE	ASEQ	ADaMIG Required
ADAE	AVAL	ADaMIG Required
ADAE	AVALC	ADaMIG Required
ADAE	PARAM	ADaMIG Required
ADAE	STUDYID	ADaMIG Required
ADAE	TRTA	ADaMIG Required
ADAE	TRTEMFL	Metadata
ADAE	TRTP	ADaMIG Required
ADAE	USUBJID	Metadata

ADSL	AGE	Metadata
ADSL	AGEGR1	Metadata
ADSL	AGEU	ADaMIG Required
ADSL	ARM	ADaMIG Required
ADSL	ETHNIC	Metadata
ADSL	FASFL	ADaMIG Required
ADSL	HEIGHTBL	Metadata
ADSL	ITTFL	Metadata
ADSL	RACE	Metadata
ADSL	SAFFL	Metadata
ADSL	SEX	Metadata
ADSL	SITEID	ADaMIG Required
ADSL	STUDYID	ADaMIG Required
ADSL	SUBJID	ADaMIG Required
ADSL	TRT01A	Metadata
ADSL	TRT01P	ADaMIG Required
ADSL	USUBJID	Metadata

ADVS	ADT	ADaMIG Required
ADVS	ADY	ADaMIG Required
ADVS	ANL01FL	Metadata
ADVS	ANLzzFL	ADaMIG Required
ADVS	ASEQ	ADaMIG Required
ADVS	AVAL	Metadata
ADVS	AVALC	ADaMIG Required
ADVS	AVISIT	Metadata
ADVS	CHG	Metadata
ADVS	PARAM	ADaMIG Required
ADVS	PARAMCD	Metadata
ADVS	STUDYID	ADaMIG Required
ADVS	TRTA	ADaMIG Required
ADVS	TRTP	ADaMIG Required
ADVS	USUBJID	ADaMIG Required

Use Case: Step 4

- **Prompt:**
  - Now, I'm going to upload the SDTM IG. Please try to understand the sources of these ADaM variables from the SDTM.

- **Results:**

1. ADAE (Adverse Events) Dataset

ADaM Variable	Likely SDTM Source	SDTM Domain
`STUDYID`	`STUDYID`	DM (Demographics)
`USUBJID`	`USUBJID`	DM
`TRTEMFL`	Derived using `EXSTDTC` or `RFXSTDTC`	EX or DM
`AEREL`	`AEREL`	AE (Adverse Events)
`AEDECOD`	`AEDECOD`	AE
`AESER`	`AESER`	AE
`AESEV`	`AESEV`	AE
`AESTDTC`	`AESTDTC`	AE
`AEENDTC`	`AEENDTC`	AE
`TRTA`	Derived from `EXTRT` or `TA` (Trial Arms)	EX / TA
`ADT`	Derived from `AESTDTC`	AE
`ADY`	Derived using `AESTDTC` relative to `RFSTDTC`	AE / DM

Use Case: Step 5

- **Prompt:**
  - Now that you have all of this information, can you please write a specs document for the ADaM datasets and their variables that you put together (All Variables Per Dataset (ARS + ADaMIG)). I need it to be in the same format at this file that I am uploading, and all of the sheets also need to be filled out. Please make sure to write methods for deriving the ADaM variables from SDTM.

- **Results:**
  - It filled out the ADaM specs! … poorly.

Dataset	Description	Class	Structure	Purpose	Key Variables	Repeating	Reference Data
ADSL	Subject-Level Analysis Dataset	One record per subject	STUDYID, USUBJID	ADSL	Contains subject-level demographics and treatment information		
ADAE	Adverse Events Analysis Dataset	One record per AE per subject	STUDYID, USUBJID, AEDECOD, AESTDTC	OCCDS	Contains records of adverse events for analysis		
ADVS	Vital Signs Analysis Dataset	One record per parameter per visit p	STUDYID, USUBJID, PARAMCD, AVISITN	BDS	Contains records of vital signs measurements		

Order	Dataset	Variable	Label	Data Type	Leng	Significant Dig	Format	Mandatory	Codelist	Origin	Pages	Method	Predecessor	Role	Comment
ADSL	AGE	Age	Numeric	Analysis	Required		SDTM.DM.AG	Direct from SDTM.DM.AGE							
ADSL	AGEGR1	Age Group 1	Character	Grouping	Derived			Derived based on AGE							
ADSL	AGEU	Age Units	Character	Supportive	Required		SDTM.DM.AG	Direct from SDTM.DM.AGEU							
ADSL	ETHNIC	Ethnicity	Character	Grouping	Required		SDTM.DM.ETH	Direct from SDTM.DM.ETHNIC							
ADSL	HEIGHT	Height (cm)	Numeric	Supportive	Permissible		Derived from	Derived from VS where VSTESTCD='HEIGHT'							

Use Case: Changing Specs Format

- **Pivoting to providing a simpler specs template:**

Order of Variables	ADaM Dataset	ADaM Variable	ADaM Variable Label	Data Type

Mandatory (Yes/No)	SDTM Predecessor Domain(s)	SDTM Predecessor Variable(s)	Description

Note: I then asked LLM to also write a Methods column for variables it labeled "Derived".

Use Case: Results

Order of Variables	ADaM Dataset	ADaM Variable	ADaM Variable Label	Data Type	Mandatory (Yes/No)	SDTM Predecessor Domain(s)	SDTM Predecessor Variable(s)	Description	Methods
1	ADSL	STUDYID	Study Identifier	Char	Yes	DM	STUDYID	Unique identifier for a study	
2	ADSL	USUBJID	Unique Subject Identifier	Char	Yes	DM	USUBJID	Unique subject identifier across studies	
3	ADSL	SUBJID	Subject Identifier for the Study	Char	Yes	DM	SUBJID	Subject identifier within a study	
4	ADSL	SITEID	Study Site Identifier	Char	Yes	DM	SITEID	Identifier for the study site	
5	ADSL	AGE	Age	Num	Yes	DM	AGE	Age at study start	
6	ADSL	AGEU	Age Units	Char	Yes	DM	AGEU	Units of age	
7	ADSL	SEX	Sex	Char	Yes	DM	SEX	Sex of the subject	
8	ADSL	RACE	Race	Char	Yes	DM	RACE	Race of the subject	
9	ADSL	ARM	Description of Planned Arm	Char	Yes	DM	ARM	Planned arm subject was assigned to	
10	ADSL	TRT01P	Planned Treatment for Period 01	Char	Yes	TA	ARMCD	Planned treatment for period 01 from trial design	
11	ADSL	SAFFL	Safety Population Flag	Char	Yes	Derived		Flag for inclusion in safety population	Set to 'Y' if subject received at least one dose (e.g., EX.EXDOSE > 0)
12	ADSL	ITTFL	Intent-To-Treat Flag	Char	Yes	Derived		Flag for inclusion in ITT population	Set to 'Y' if subject was randomized
13	ADSL	FASFL	Full Analysis Set Flag	Char	Yes	Derived		Flag for inclusion in full analysis set	Set to 'Y' based on study definition of full analysis set (e.g., treated + baseline data)
14	ADSL	RFSTDTC	Subject Reference Start Date/Time	Char	Yes	DM	RFSTDTC	Reference start date for the subject	
15	ADSL	RFENDTC	Subject Reference End Date/Time	Char	Yes	DM	RFENDTC	Reference end date for the subject	
1	ADAE	STUDYID	Study Identifier	Char	Yes	DM	STUDYID	Unique identifier for a study	
2	ADAE	USUBJID	Unique Subject Identifier	Char	Yes	DM	USUBJID	Unique subject identifier across studies	
3	ADAE	TRTEMFL	Treatment Emergent Flag	Char	Yes	Derived		Flag for adverse events occurring after treatment start	Set to 'Y' if AESTDTC is on or after RFSTDTC; otherwise null
4	ADAE	AEREL	Causality	Char	Yes	AE	AEREL	Causality assessment of adverse event	
5	ADAE	AEDECOD	Dictionary-Derived Term	Char	Yes	AE	AEDECOD	Standardized AE term from dictionary	
6	ADAE	AESER	Serious Event	Char	Yes	AE	AESER	Indicates if the AE is serious	
7	ADAE	AESEV	Severity/Intensity	Char	Yes	AE	AESEV	Severity of the adverse event	
8	ADAE	AESTDTC	Start Date/Time of AE	Char	Yes	AE	AESTDTC	Start date/time of adverse event	
9	ADAE	AEENDTC	End Date/Time of AE	Char	Yes	AE	AEENDTC	End date/time of adverse event	
10	ADAE	TRTA	Actual Treatment	Char	Yes	EX	EXTRT	Actual treatment administered	
11	ADAE	ADT	Analysis Date	Char	Yes	Derived	AESTDTC	Converted analysis date	Convert AESTDTC or VSDTC from character to numeric date
12	ADAE	ADY	Analysis Relative Day	Num	Yes	Derived	AESTDTC, RFSTDTC	Relative day from reference start date	ADY = ADT - RFSTDTC (in days); ADY is relative to RFSTDTC; Day 1 = first dose
1	ADVS	STUDYID	Study Identifier	Char	Yes	DM	STUDYID	Unique identifier for a study	
2	ADVS	USUBJID	Unique Subject Identifier	Char	Yes	DM	USUBJID	Unique subject identifier across studies	
3	ADVS	PARAM	Parameter	Char	Yes	VS	VSTESTCD/VSTEST	Analysis parameter name	
4	ADVS	AVAL	Analysis Value	Num	Yes	VS	VSSTRESN	Standardized result value	
5	ADVS	AVISIT	Analysis Visit	Char	Yes	VS	VISIT	Visit at which data was collected	
6	ADVS	ADY	Analysis Relative Day	Num	Yes	Derived	VSDTC, RFSTDTC	Relative day from reference start date	ADY = ADT - RFSTDTC (in days); ADY is relative to RFSTDTC; Day 1 = first dose
7	ADVS	DTYPE	Derivation Type	Char	No	Derived		Type of derivation (e.g., LOCF, WOCF)	Populated as 'LOCF', 'WOCF', etc. based on imputation rules for missing data
8	ADVS	ABLFL	Baseline Flag	Char	No	Derived		Indicates baseline value	Set to 'Y' for the last non-missing value prior to first dose (ADT < RFSTDTC)
9	ADVS	ANL01FL	Analysis Flag 01	Char	No	Derived		Flag to identify record used in analysis	Flag set to 'Y' for record used in primary analysis, per SAP rules
10	ADVS	TRTP	Planned Treatment	Char	Yes	TA	ARMCD	Planned treatment for analysis	
11	ADVS	TRTA	Actual Treatment	Char	Yes	EX	EXTRT	Actual treatment administered	

Use Case: Results

- **Derived:**

11	ADSL	SAFFL	Safety Population Flag	Set to 'Y' if subject received at least one dose (e.g., EX.EXDOSE > 0)
12	ADSL	ITTFL	Intent-To-Treat Flag	Set to 'Y' if subject was randomized
13	ADSL	FASFL	Full Analysis Set Flag	Set to 'Y' based on study definition of full analysis set (e.g., treated + baseline data)

- **Predecessor:**

ADVS	PARAM	Parameter	Char	Yes	VS	VSTESTCD/VSTEST
ADVS	AVAL	Analysis Value	Num	Yes	VS	VSSTRESN
ADVS	AVISIT	Analysis Visit	Char	Yes	VS	VISIT

Learnings

- Generative AI can take inputs from ARS metadata, Protocol, SAP, and ADaM/SDTM Implementation Guide to build reliable and detailed ADaM specifications.
- If Generative AI has a logical model to depend on like the ARS metadata (structured and machine-readable foundation) then the results that it can generate with repetitive prompting will get better.
- This kind of automation helps reduce manual effort, cuts down errors, and improves traceability across the entire clinical workflow.
- Accuracy & Reliability
  - While AI can generate plausible outputs, they must be carefully reviewed
  - AI can support, but not fully replace the clinical, statistical, and regulatory knowledge needed for robust ADaM specs development.
  - Needs Human-in-the-loop Validation

What's Next for Us

- Living Specifications
  - With Gen-AI, ADaM specs don't have to be static - they can change and adapt in real time as protocols and SAPs evolve. This is especially helpful for agile or adaptive trials. Gen-AI needs rules and Models and ARS metadata will do that.
- Standardization at Scale
  - We can start using ARS metadata-driven templates and rules across studies and teams, which helps ensure consistent application of CDISC standards organization-wide.
- Regulatory Readiness
  - The outputs generated by AI - like Define-XML or ODM - can be aligned early with submission needs, making it easier to get ready for regulatory reviews and reducing last-minute surprises.
- Collaboration & Governance
  - A metadata-first approach also supports better collaboration between functions -statisticians, programmers, data managers - and makes it easier to track derivations and ensure compliance through automated traceability.

Questions?

Anna Yaggi
TFL Designer Technical SME and
Statistical Programmer
ayaggi@clymbclinical.com

Shivani Gupta
Director of Stats Programming
sgupta@clymbclinical.com

EMAIL
info@clymbclinical. com

WEBSITE
www.clymbclinical.com

PHONE
781-692-3613

References

"Analysis Results Standard | CDISC." *Cdisc.org*, 19 Apr. 2024, www.cdisc.org/standards/foundational/analysis-results-standard. Accessed 5 June 2025.
"ARS Trainings." *Cdisc.org*, 2025, learnstore.cdisc.org/catalog?pagename=ARS. Accessed 5 June 2025.
"ARS-UG Sections - Analysis Results Standard Document - Wiki." *Cdisc.org*, 2024, wiki.cdisc.org/display/ARSP/ARS-UG+Sections. Accessed 5 June 2025.
Ciotti, Melanie, et al. *Conclusions Figure 1b: Observed Data and SDTM Predictions for RACE (Multiple Selections)*
Figure 1a: Observed Data and SDTM Predictions for RACE (Single Selection) Figure 2: Sample CDISC Code List for Variable RACE
Figure 3: SDTM Annotated Case Report Forms Generated by Artificial Intelligence (RACE as Single Selection on Left;
RACE as Multiple Selections on Right) Automating SDTM for Clinical Trial Data Submissions Using Artificial Intelligence and Machine Learning.
Ebi, Kunihito, et al. *Integrating Digital Data Flow with Generative AI for Enhanced Clinical Trial Automation*.
Kim, Minseok . *Application of AI in Clinical Trials*.
Leroy, Bess, and Richard Marshall. *Getting Started with the New CDISC Analysis Results Standard*.
"The Complete History of the ChatGPT Logo - Hatchwise." *Hatchwise*, 2 June 2024, hatchwise.com/resources/the-complete-history-of-the-chatgpt-logoOffice.
"FDA Launches Agency-Wide AI Tool to Optimize Performance for the American People." *U.S. Food and Drug Administration*, 2025,
www.fda.gov/news-events/press-announcements/fda-launches-agency-wide-ai-tool-optimize-performance-american-people.
. Accessed 5 June 2025.