

LLM Powered Clinical Data Flow with Traceability - From Siloed Pipelines to a Semantic Fabric

Pankaj Attri, Solutions Architect, SAS

Matt Becker, Life Sciences Strategic Advisor, SAS



Bio



Pankaj Attri
Pre-Sales Solutions Architect
SAS, Commercial Healthcare & Life Sciences

Pankaj Attri is an AI/ML Solutions Architect focused on leveraging advanced data science and artificial intelligence to support customers in the Life Sciences and Healthcare industries at SAS. He partners with these organizations to design and implement state-of-the-art custom solutions that accelerate innovation across research, development, and clinical operations.

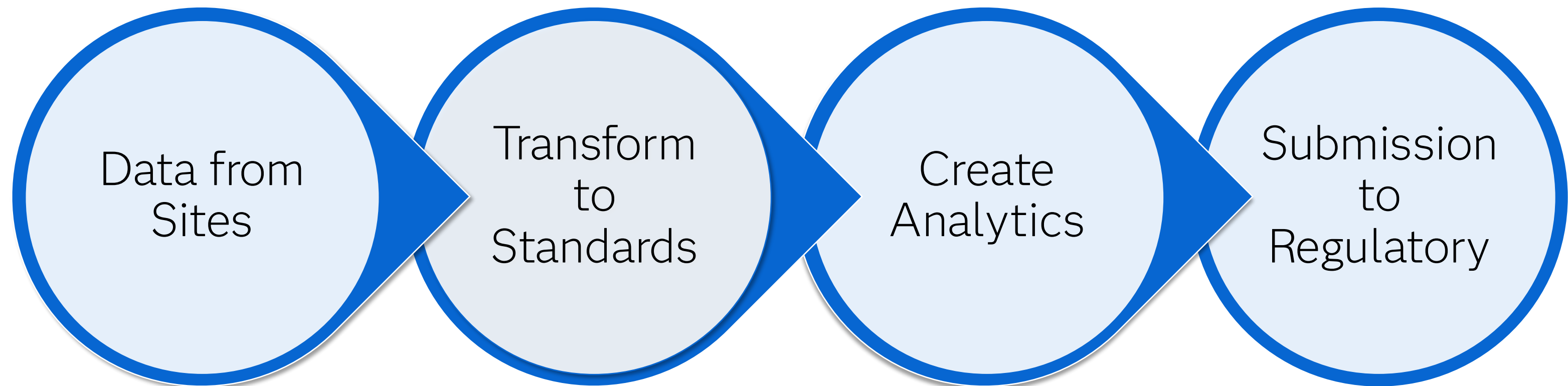
Pankaj brings together a background in Chemical Engineering (B.Tech) from IIT Bombay, India and a Master's in Computer Science from NC State University, USA.



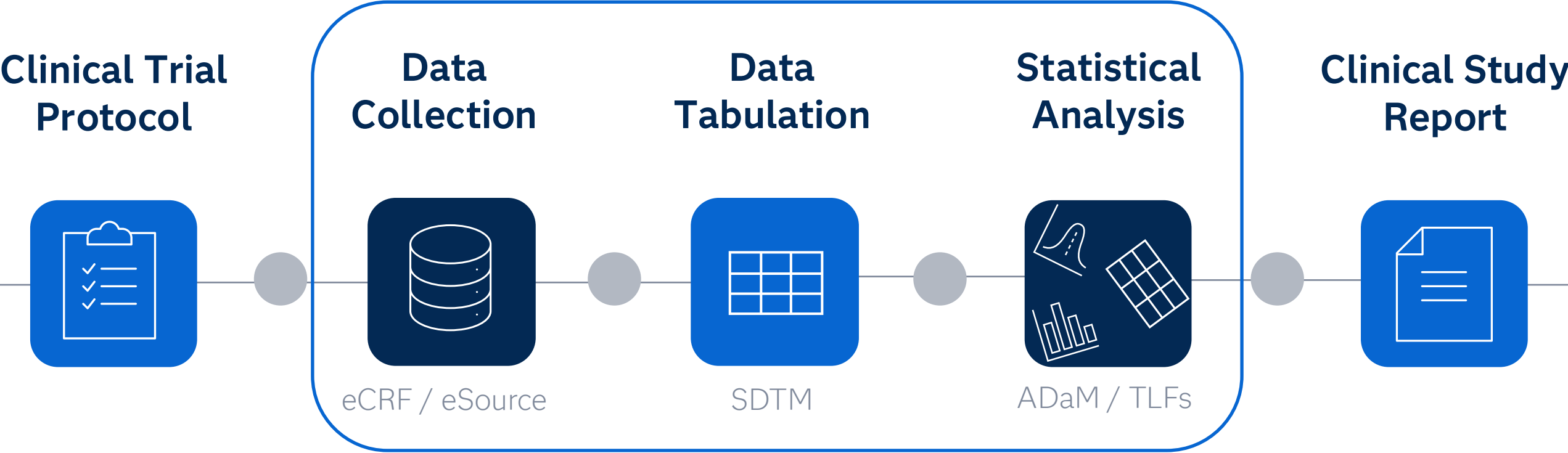
Matt Becker
Life Sciences Strategic Advisor
SAS, Healthcare and Life Sciences Industry Solutions

Matt Becker serves as a Life Science Strategic Advisor within the SAS Health Care and Life Sciences Industry Solutions division. Mr. Becker brings to this role over 35 years of experience in the life sciences, including more than 13 years with SAS. His areas of focus include next-generation clinical trials, clinical analytic platforms, data management, advanced analytics, and deployment options within the Life Science sector.

Clinical Trial Biometrics High Level Data Flow



End to End Clinical Data Flow



Raw data to SDTM format

SDTM to ADaM

ADaM Creation & Validation

Creation of define.xml

Creation of TLFs

Challenges in Current State of Clinical Trial Biometrics

Clinical trials are increasingly complex

- Tighter regulations
- Intricate data sources
- Traceability requirements

Many processes still rely on manual mapping definitions / programming

- Even with target metadata standards

Manual programming can be repetitive, inefficient, and prone to errors

- Double Programming
- Repetitive coding tasks
- Impacts submission deadlines

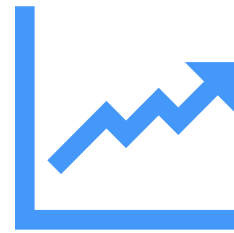
The Need for Automation



Manual data mapping
and coding

Time-consuming

Resource-intensive



Automation can
improve accuracy and
reduce burden



As trials expand globally, the
demand for streamlined
processes rises

AI Agents and LLMs: Impact on Clinical Programming

- Agentic AI and LLMs augment clinical programmers across the clinical data flow lifecycle.
- Streamline validation cycles by accelerating QC and reducing iterative review overhead.
- Ensure synchronization between data and metadata from ingestion through analysis and outputs.

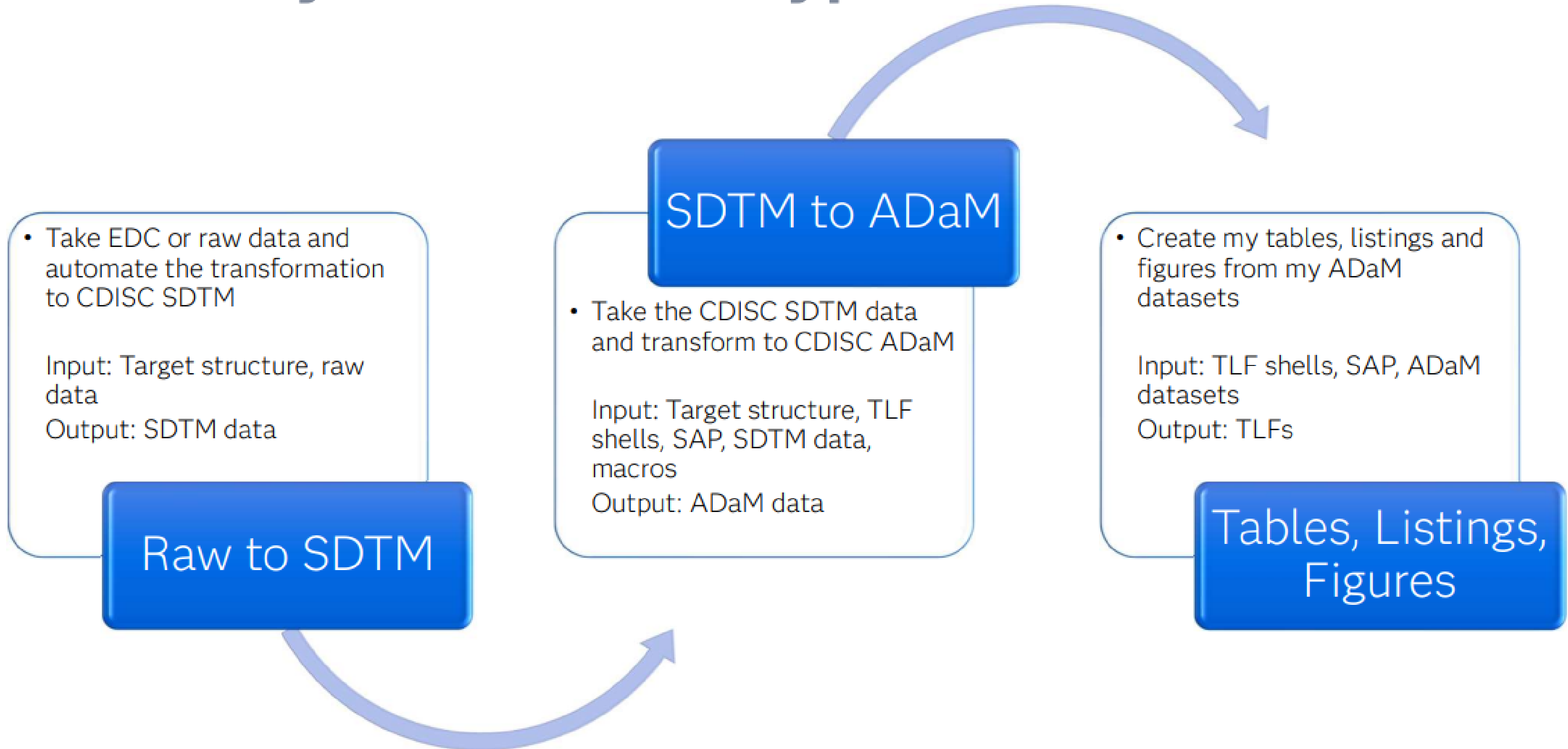
Goals of the Session

- Present practical strategies for using LLMs to automate mapping and code generation while adhering to CDISC standards to shorten time-to-submission.
- Show how AI enhances end-to-end traceability and auditability.
- Present a Minimally Viable Prototype for integrating AI into clinical workflows to mitigate regulatory risks and accelerate FDA submission timelines.

Clinical Data Flow (CDF)

Minimum Viable Prototype

Minimally Viable Prototype - Goals



Features

1

Data Mapping

- Auto generate data transformation rules. (Raw -> SDTM -> ADaM)
- Investigate documents such as Study Protocols, SAP, TLF Shells to generate mapping specifications.

2

Code Generation

- SAS Code generation with error feedback loop.
- Ensure Cyclic dependencies are taken care of when generating datasets.

3

Traceability

- Lineage view to track source to target data transformation.
- Source data coverage, missed variables reports.

1

CDF App – Mapping Rules Generation

Welcome, sinpan

Clinical Data Flow AI Assistant

Project Name: Study-ONCO-IMMUNO-25

LLM Outputs

Estimated Cost: \$0.00

Clinical Data Transformation Steps

Convert Raw Data to SDTM standards

Start here as the first step. Press Begin next to 'Identify' to get SDTM datasets, variables, and mapping logic. Once satisfied, click Start next to 'Generate' to produce SDTM Datasets & SAS code.

Identify SDTM Datasets:

Begin

Generate SDTM Datasets:

Start

State:

In-Process

Extract TLF shells logic

After indicating valid SDTM and ADaM IG versions and SDTM domains upload the TLF shells document. After uploading, press Start to extract TLF shell details. When content is satisfactory, select "Approved" in status to proceed.

TLF Shells:

Choose File

No file chosen

Extract

State:

In-Process

Identify and Generate ADaM Datasets

Upload the SAP document and press Begin next to

Agent Outputs

SDTM Datasets

SDTM Code

ADaM Datasets

ADaM Code

TLF Logic

TLF Code

↶

✎

⬇

↶

Domain	Variable	Type	Description	Mapping Rule	Data Source	Confidence	Status
No data available							

Clinical Data Flow Steps

No SDTM datasets available

1

CDF App –Mapping Rules Generation - SDTM

Agent Outputs

SDTM Datasets

SDTM Code

ADaM Datasets

ADaM Code

TLF Logic

TLF Code

Domain	Variable	Type	Description	Mapping Rule	Data Source	Confidence	Status
AE	AEACN	Expected	Describes changes to study treatment as a result of the event.	Map from ADVERSE_EVENTS.ACTION_TAKEN; if empty, leave null. If sponsor codelist requires, map site terms to ICH E2B (e.g., 'DOSE NOT CHANGED', 'DRUG WITHDRAWN').	ADVERSE_EVENTS.ACTION_TAKEN	Medium	
AE	AEBDSYCD	Expected	Code for the body system or organ class used by the sponsor.	Not available in raw data. Set to null.		Low	
AE	AEBODSYS	Expected	Body system or organ class used by the sponsor.	Map directly from ADVERSE_EVENTS.AE_BODSYS.	ADVERSE_EVENTS.AE_BODSYS	High	
AE	AEDECOD	Required	Dictionary-derived text description of AETERM (Preferred Term).	Map directly from ADVERSE_EVENTS.AE_DECOD.	ADVERSE_EVENTS.AE_DECOD	High	
AE	AEENDTC	Expected	End date/time of the adverse event in ISO 8601 character format.	Map directly from ADVERSE_EVENTS.END_DATE; may be blank if ongoing or not recorded.	ADVERSE_EVENTS.END_DATE	High	
AE	AEENDY	Permissible	Study day of end of adverse event relative to RFSTDTC.	Derive as AEENDY = (AEENDTC - RFSTDTC) + 1 where RFSTDTC = PATIENT_INFORMATION.FIRST_TREATMENT_DATE; null if AEENDTC missing.	ADVERSE_EVENTS.END_DATE, PATIENT_INFORMATION.FIRST_TREATMENT_DATE	Medium	
AE	AEHLGT	Expected	High Level Group Term.	Map directly from ADVERSE_EVENTS.AE_HLGT.	ADVERSE_EVENTS.AE_HLGT	High	
AE	AEHLGTC	Expected	Code for the High Level Group Term.	Derive numeric code by extracting digits from ADVERSE_EVENTS.AE_HLGT values like 'HLGT_0192' -> 192.	ADVERSE_EVENTS.AE_HLGT	Medium	

1

CDF App – Mapping Rules Generation - SDTM

Agent Outputs

SDTM Datasets

SDTM Code

ADaM Datasets

ADaM Code

TLF Logic

TLF Code

Domain	Variable	Type	Description	Mapping Rule	Data Source	Confidence	Status	Action
SE	USUBJID	Required	Source.		USUBJID	High		Delete
SUPPAE	IDVAR	Expected	Identifying Variable	ASSIGN_CONSTANT [AESEQ]		High		Delete
SUPPAE	IDVARVAL	Expected	Identifying Variable Value	DIRECT_COPY [AE.AESEQ]	AE.AESEQ	High		Delete
SUPPAE	QEVAL	Expected	Evaluator	ASSIGN_NULL []		High		Delete
SUPPAE	QLABEL	Required	Qualifier Variable Label	TRANSPOSE_LABEL [Treatment Emergent Flag]	ADVERSE_EVENTS.TREATMENT_EMERGENT_FL	High		Delete
SUPPAE	QNAM	Required	Qualifier Variable Name	TRANSPOSE_ID [TRTEMERG<=ADVERSE_EVENTS.TREATMENT_EMERGENT_FL]	ADVERSE_EVENTS.TREATMENT_EMERGENT_FL	High		Delete
SUPPAE	QORIG	Required	Origin	ASSIGN_CONSTANT [Derived]		Medium		Delete
SUPPAE	QVAL	Required	Qualifier Variable Value	TRANSPOSE_VALUE [ADVERSE_EVENTS.TREATMENT_EMERGENT_FL]	ADVERSE_EVENTS.TREATMENT_EMERGENT_FL	High		Delete

1

CDF App – Mapping Rules Generation - ADaM

Project Name: Xanomeline LY-246708

Estimated Cost: \$4.45

Clinical Data Transformation Steps

Extract TLF shells logic

After indicating valid SDTM and ADaM IG versions and SDTM domains upload the TLF shells document. After uploading, press Start to extract TLF shell details. When content is satisfactory, select "Approved" in status to proceed.

TLF Shells:

Choose File

 No file chosen

Extract

State:

Approved

Identify and Generate ADaM Datasets

Upload the SAP document and press Begin next to 'Identify' to get ADAM datasets, variables, and mapping logic. Once satisfied, click Start next to 'Generate' to produce ADaM Datasets & SAS code.

SAP:

Choose File

 CDISCPLOT_SAP.pdf ☐ No SAP available

Identify ADaM Datasets:

Begin

Macros File:

Choose File

 Pre_DefinedMacros.sas

Generate ADaM Datasets:

Start

State:

In-Process

Agent Outputs

SDTM Datasets

SDTM Code

ADaM Datasets

ADaM Code

TLF Logic

TLF Code

Dataset Name

Variable

Type

Description

Mapping Rule

Source

Data Source

Confidence

Status

ADAE	ADURN	permissible	Analysis Duration (N)	Derived from ASTDT/ASTDTM and AENDT/AENDTM (e.g., days	ADaM IG	AE.AESTD TC,AE.AE ENDTC	High	
ADAE	ADURU	conditional	Analysis Duration Units	Units for ADURN (e.g., 'DAYS').	ADaM IG	AE.AESTD TC,AE.AE ENDTC	High	
ADAE	AEACN	conditional	Action Taken with Study Treatment	Direct mapping from SDTM AE.AEACN.	ADaM IG	AE.AEAC N	High	
ADAE	AEBDSYCI	required	Body System or Organ Class Code	Direct mapping from SDTM AE.AEBDSYCD or supplemental qualifier.	ADaM IG	AE.AEBD SYCD	High	
ADAE	AEBODSY:	required	Body System or Organ Class	Direct mapping from SDTM AE.AEBODSYS.	ADaM IG	AE.AEBO DSYS	High	
ADAE	AECAT	permissible	Category	Direct mapping from SDTM AE.AECAT.	ADaM IG	AE.AECAT	High	
ADAE	AEDECOD	required	Dictionary-Derived Term	Direct mapping from SDTM AE.AEDECOD (MedDRA Preferred Term).	ADaM IG	AE.AEDE COD	High	
ADAE	AEDUR	conditional	Duration of XX	Direct mapping of collected duration from SDTM AE.AEDUR.	ADaM IG	AE.AEDU R	High	

Copyright © SAS Institute Inc. All rights reserved.

CDF App – Code Generation

The screenshot displays the CDF App interface, specifically the SDTM Domian Tab. The interface includes a top navigation bar with tabs for SDTM Datasets, SDTM Code (highlighted), ADaM Datasets, ADaM Code, TLF Logic, and TLF Code. Below this, a row of domain-specific tabs is shown, with 'QS' (Questionnaires) highlighted. A 'Code' tab is also visible. The 'Last Run Log' and 'Last Run Output' sections are present. The 'Run Code' button is highlighted, and the 'Last Run Status' is 'Success'. The generated SAS code is displayed in a text area, showing the program structure and the SQL query used to select the STUDYID.

Agent Outputs

SDTM Domian Tab

SDTM Datasets | **SDTM Code** | ADaM Datasets | ADaM Code | TLF Logic | TLF Code

QS | SE | SC | DM | MH | TV | LB | AG | TE | EX | CM | AE | TS | TI | RELREC | SV | TA | VS | DS

Code | Last Run Log | Last Run Output

Run Code | Save Code | Last Run Status: **Success** ✓

Edit Code
Re-generate Code
Manually Run Code

Generated and Run SAS Code, Logs, HTML Outputs

```
*-----;
* PROGRAM:      create_sdtm_qs.sas                               ;;
* DESCRIPTION:   Creates the SDTM Questionnaires (QS) domain from raw data.  ;;
*               ;;
* INPUT:         RAW.QUESTIONNAIRES, RAW.VISIT_SCHEDULE, RAW.PATIENT_INFORMATION, ;;
*               RAW.PROTOCOL_DEFINITIONS                               ;;
* OUTPUT:        SDTM.QS                                             ;;
*               ;;
* AUTHOR:        Clinical Data Programmer AI                       ;;
* DATE:          16 November 2025                                   ;;
*               ;;
* NOTES:         This program adheres to the provided SDTM specifications and mapping rules. ;;
*-----;

options nocenter ls=132 ps=64;

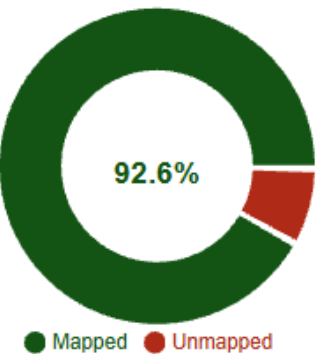
*-----;
* Define LIBNAMEs for source Raw data and target SDTM datasets.      ;;
*-----;
libname RAW  "/nfsshare/sashls2/data/sinpan/Clinical_Data_Flow_AutomationwithLLMs/CDISCPILLOT_DEMO/raw-data";
libname SDTM "/nfsshare/sashls2/data/sinpan/Clinical_Data_Flow_AutomationwithLLMs/CDISCPILLOT_DEMO/tabulations-sdtm";

*-----;
* Obtain STUDYID constant from PROTOCOL_DEFINITIONS.STUDY_ID using a robust single-value selection.;;
*-----;
%let STUDYID=;
proc sql noprint;
  select strip(min(study_id))
    into :STUDYID
```

3 CDF App – Traceability and Data Coverage Reports

Data Mapping Analysis

Overall Raw Data Coverage



Total Variables: 149

Raw Dataset Mapping Details

Dataset	Total Vars	Mapped	Unmapped
ADVERSE_EVENTS	23	23	0
CONCOMITANT_MEDICATIONS	11	11	0
DRUG_ADMINISTRATION	9	9	0
INCLUSION_EXCLUSION	3	0	3
LABORATORY_RESULTS	18	18	0
MEDICAL_HISTORY	6	6	0
PATIENT_DISPOSITION	8	7	1
PATIENT_INFORMATION	25	25	0
PROTOCOL_DEFINITIONS	4	4	0
QUESTIONNAIRES	12	12	0
STUDY_SCHEDULE	10	4	6
VISIT_SCHEDULE	9	8	1
VITAL_SIGNS	11	11	0

SDTM Unmapped Variables

Domain Variables (Rule: Not in Raw)	
AE	AEACNDEV, AEACNOth, AEBDSYCD, AELLTCD, AELOC, AEPATT, AEPTCD, AESOCCD
CM	CMDOSFRM
DM	ACTARMUD
EX	EXAD, EXCAT, EXCDID, TACTOPP

Raw to Data Standards Lineage

Filter Raw Datasets

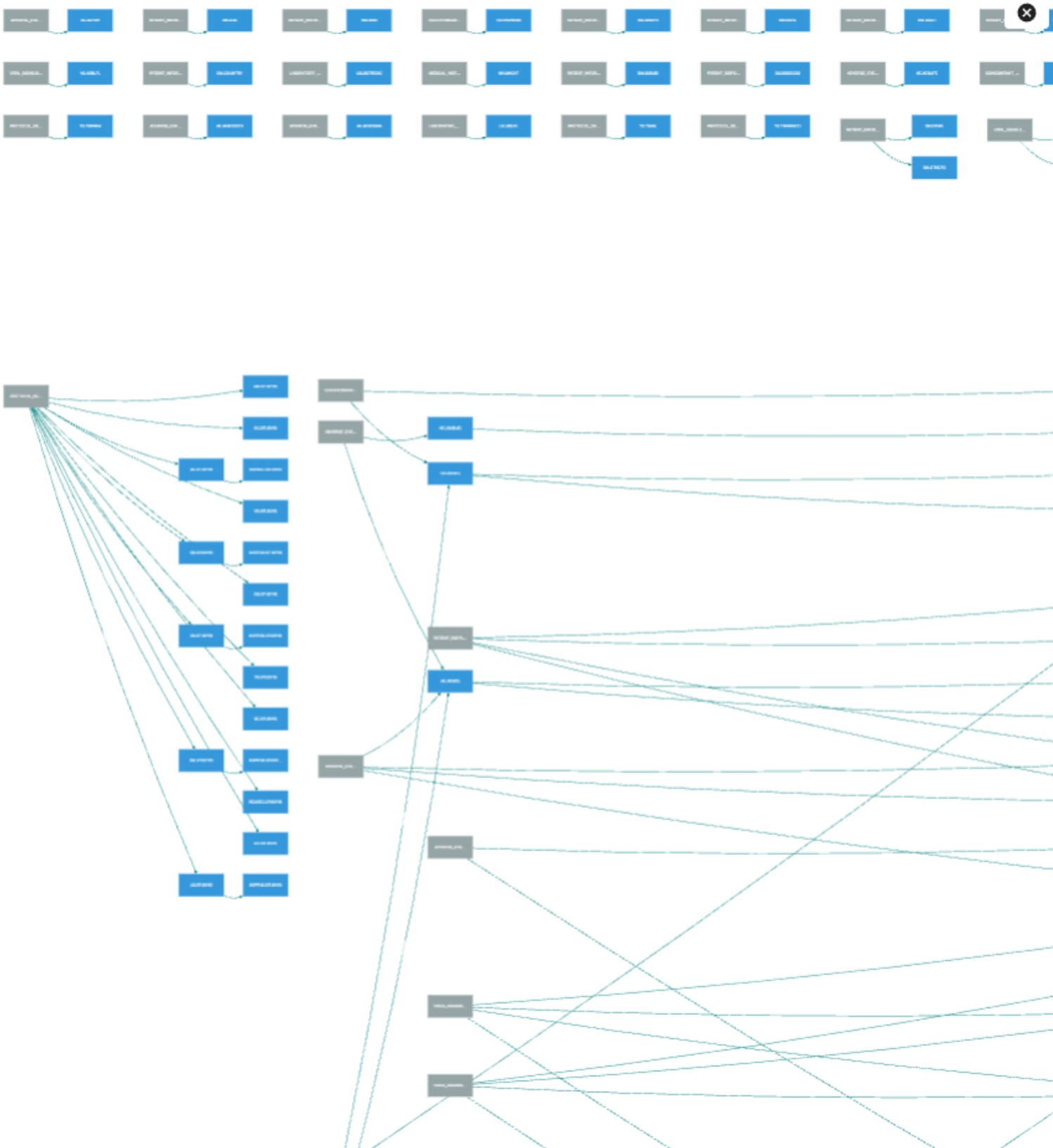
Clear Filter (0 selected)

Filter SDTM Datasets

Clear Filter (0 selected)

Graph Legends

- Raw Dataset
- SDTM Dataset



3

CDF App – Traceability and Data Coverage Reports

Details

Study Name: Safety and Efficacy of the Xanomeline Transdermal Therapeutic System (TTS)

Study Objectives:

Primary Objective 1 (Efficacy): To determine if there is a statistically significant relationship between the change in both the ADAS-Cog (11) and CIBIC+ scores, and drug dose. (SAP Section 3.1.1: 'To determine if there is a statistically significant relationship ... between the change in both the ADAS-Cog (11) and CIBIC+ scores, and drug dose')

Primary Objective 2 (Safety): To document the safety profile of the xanomeline TTS. (SAP Section 3.1.1: 'To document the safety profile of the xanomeline TTS.')

Secondary Objective (Efficacy): To assess the dose-dependent improvement in behavior, indicated by improved scores on the Revised Neuropsychiatric Inventory (NPI-X). (SAP Section 3.1.2: 'To assess the dose-dependent improvement in behavior.')

Primary Endpoint 1: Alzheimer's Disease Assessment Scale - Cognitive Subscale, total of 11 items [ADAS-Cog (11)] at Week 24. (SAP Section 3.2.1)

Primary Endpoint 2: Video-referenced Clinician's Interview-based Impression of Change (CIBIC+) at Week 24. (SAP Section 3.2.1)

Secondary Endpoint 1: ADAS-Cog (11) at Weeks 8 and 16. (SAP Section 3.2.2)

Secondary Endpoint 2: CIBIC+ at Weeks 8 and 16. (SAP Section 3.2.2)

Secondary Endpoint 3: Mean Revised Neuropsychiatric Inventory (NPI-X) from Week 4 to Week 24. (SAP Section 3.2.2)

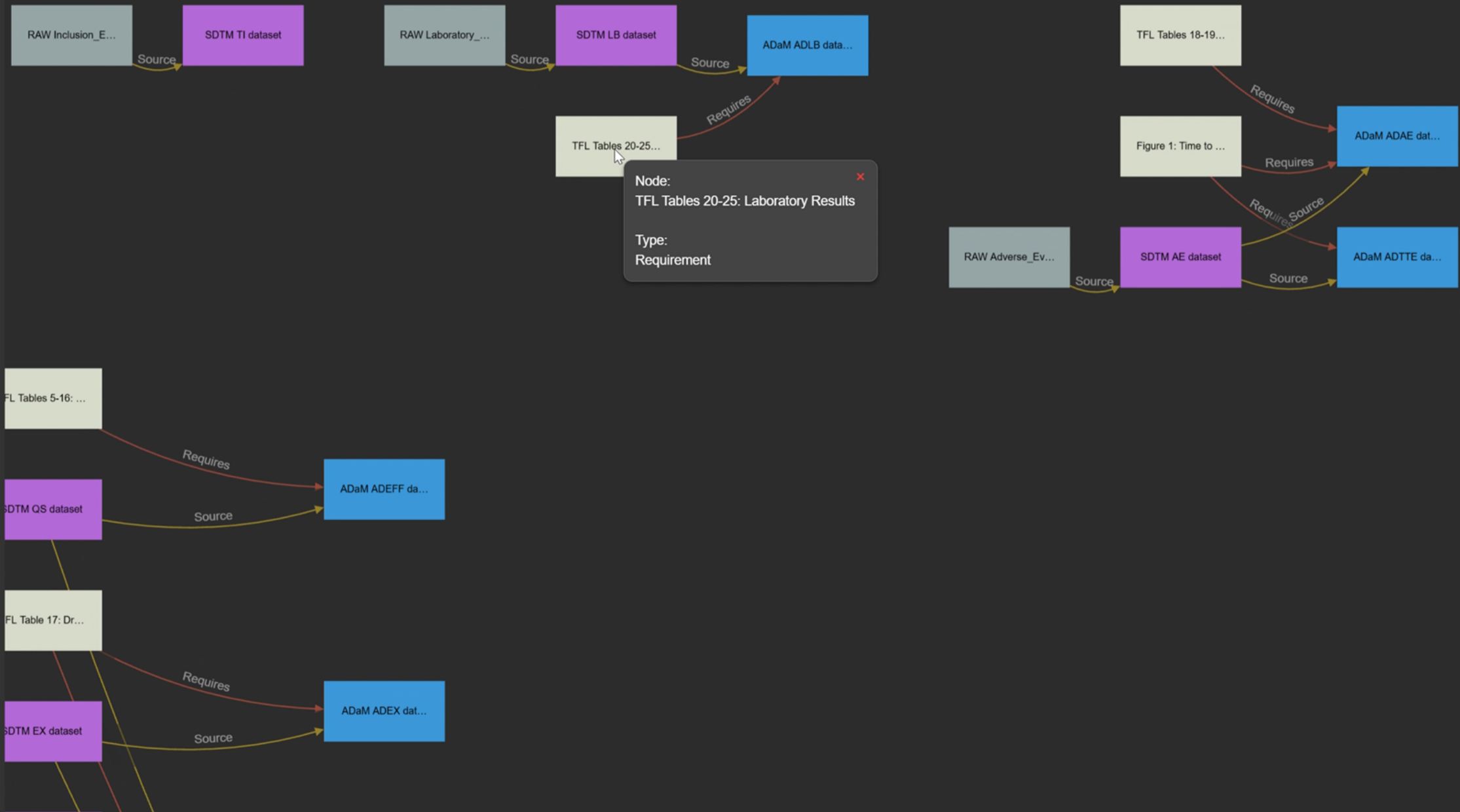
Safety Endpoints: Adverse events, vital signs (weight, blood pressure, heart rate), and laboratory evaluations. (SAP Section 3.2.2)

Graph Node Legend

- Requirement as specified in SAP or TFL
- ADaM Dataset
- SDTM Dataset
- RAW Dataset

Graph Edge Legend

- Source
- Requires



Close

Business Impact

Automated mapping & code generation - faster builds, lower cost

LLMs enhance rather than replace human programmers, transforming routines into streamlined processes.

Compress timelines: mapping, SDTM/ADaM, and TLF generation shrink from months to days/hours.

Jump-start, not turnkey: automate ~70–80% to create a robust baseline; humans finalize for submission readiness.

Reallocate expertise to high-value work (e.g., complex efficacy datasets), high-level validation, scientific analyses.

Responsible Adoption of Gen AI

Responsible Adoption

Mitigate risks associated with LLMs through secure, private models and human oversight.

Validation of AI-generated outputs is essential to ensure coding accuracy.

Create New Business processes around consumption of Gen AI technologies.

Question and Answers

For any questions contact:
Pankaj.attri@sas.com, Matt.becker@sas.com

