SDTM-based data cleaning
基于SDTM的数据清理

Ze Xu/Senior Manager of Data Insights, Beone

29Aug2025

SDTM-based data cleaning 基于SDTM的数据清理

Ze Xu/Senior Manager of Data Insights, Beone

Meet the Speakers

Ze Xu

Title: Senior Manager of Data Insights

Organization: Beone

Xiaojuan QU

Title: Associate Director of Data Management

Organization: Beone

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

Agenda

1. The status of Beone Data Cleaning
2. The Prerequisites for Data Cleaning Based on SDTM
3. BeOne SDTM Status
4. Challenges of Conducting Data Cleaning Based on SDTM
5. Future Outlook

The status of Beone Data Cleaning

The status of Beone Data Cleaning

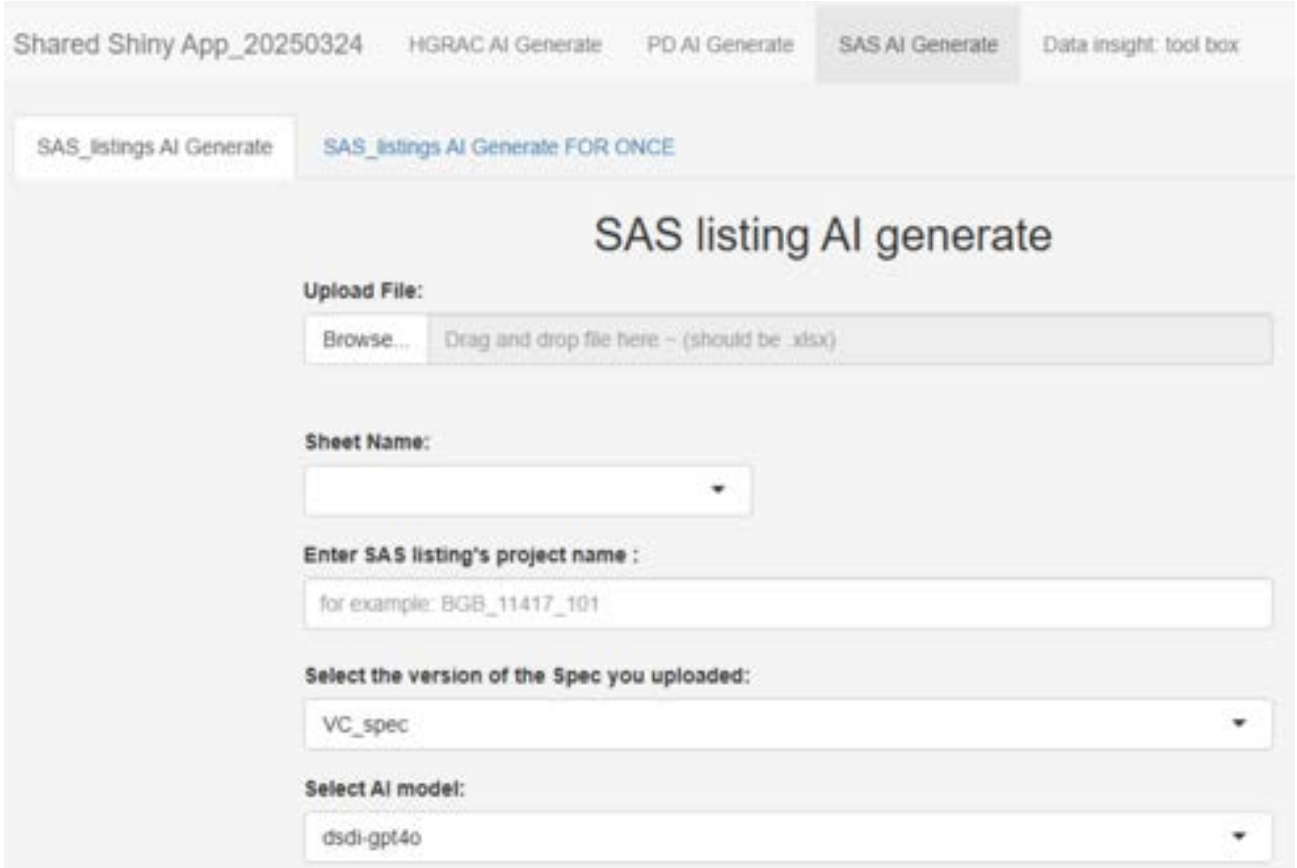
Three Steps:

1. Edit Check: System Query, Dynamic Query
2. CEDAR(Clinical Enterprise Data Acquisition Repository) system
3. Programming by SAS/R/Python

The status of Beone Data Cleaning

- Developed semi-automated tools for Edit Check.
- Standard SAS Listings can now be automatically executed in the CEDAR system and are connected to the backend of the RAVE system for automatic query dispatch.
- Study-specific SAS Listings are managed by different DMs across projects. The logical style of specifications varies, resulting in low code reuse between projects.
- AI-assisted programming is effective for simple logic, but complex logic still requires further manual review and writing.
- A UAT input system is under development. Currently, UAT is not interchangeable between studies. Multiple rounds of UAT testing are required, and UAT data entry is time-consuming, leading to long delivery times from programming to the final report.

AI-assisted programming



The Prerequisites for Data Cleaning Based on SDTM

The Prerequisites for Data Cleaning Based on SDTM

Company-Level Design: Establish a robust library of standard protocols and CRFs, with strict review processes for non-standard CRFs included in studies.

Efficient Data Transformation: Develop a rapid, standardized, high-quality coding framework/tool for converting raw data to SDTM.

Listing Reports: Generate listing reports based on SDTM that preserve key raw data information, enabling DM to promptly document queries.

Standard UAT: Implement a standard UAT process based on SDTM.

Example – CTCAE Calculate

- Lab (Local) <--> Adverse Events

- Review Local Lab that algorithmically meet the CTCAE definition for an Adverse Event and compare to reported Adverse Events/Medical History:

- 1) Lab-> AE, if Lab is flagged with CTCAE grade >= 3 and no associated AE ((AE duration = AE start date - 3 days to AE stop date) and AE (grade >= CTCAE grade)) or Ongoing MH, query to site.

- 2) AE-> Lab, if Lab AE exists and there is no related lab record with acceptable CTCAE grade flag in AE duration (AE start date - 3 days to AE stop date). Query to site
- * if AE grade = 5 and the CTCAE may = 4 or 5
- * if AE grade < 5 and the CTCAE should = AE grade
- Note: for grade consistent check, please consultant with CD if not sure.

Example – CTCAE Calculate

- To perform lab CTCAE grade calculation, these are the required input factors:
  - LBCAT, LBTEST, LBSTRESU, LBSTRESEN, NORMAL RANGE, BASELINE RESULT(only for certain parameters)

Investigations					
CTCAE Term	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**	**Grade 5**
Activated partial thromboplastin time prolonged	>ULN - 1.5 x ULN	>1.5 - 2.5 x ULN	>2.5 x ULN; bleeding	-	-
Definition: A finding based on laboratory test results in which the partial thromboplastin time is found to be greater than the control value. As a possible indicator of coagulopathy, a prolonged partial thromboplastin time (PTT) may occur in a variety of diseases and disorders, both primary and related to treatment. **Navigational Note:** -					
Alanine aminotransferase increased	>ULN - 3.0 x ULN if baseline was normal; 1.5 - 3.0 x baseline if baseline was abnormal	>3.0 - 5.0 x ULN if baseline was normal; >3.0 - 5.0 x baseline if baseline was abnormal	>5.0 - 20.0 x ULN if baseline was normal; >5.0 - 20.0 x baseline if baseline was abnormal	>20.0 x ULN if baseline was normal; >20.0 x baseline if baseline was abnormal	-
Definition: A finding based on laboratory test results that indicate an increase in the level of alanine aminotransferase (ALT or SGPT) in the blood specimen. **Navigational Note:** Also consider Hepatobiliary disorders: Hepatic failure					

Example – CTCAE Calculate

- CTCAE metadata and macros

md_toxgr_ctcae5

Filter and Sort | Query Builder | Where | Data | Describe | Graph | Analyze | Export | Send To

	LBCAT	LBTEST	SIUNIT	L1	L2	L3	L4	H1	H2	H3	H4	BLCH1	BLCH2	BLCH3	BLCH4
1	COAGULATION	Activated Partial Thromboplastin..						ULN - 1.5*ULN	1.5*ULN - 2.5*..	2.5*ULN					
2	BIOCHEMISTRY	Alanine Aminotransferase						ULN - 3*ULN	3*ULN - 5*ULN	5*ULN - 20*ULN	20*ULN	1.5*BLN - 3*BL..	3.0*BLN - 5*BL..	5.0*BLN - 20.0..	20.0*BLN
3	BIOCHEMISTRY	Albumin	g/L	LLN - 30	30 - 20	20									
4	BIOCHEMISTRY	Alkaline Phosphatase						ULN - 2.5*ULN	2.5*ULN - 5*UL..	5*ULN - 20*ULN	20*ULN	2.0*BLN - 2.5*..	2.5*BLN - 5.0*..	5*BLN - 20*BLN	20*BLN
5	BIOCHEMISTRY	Aspartate Aminotransferase						ULN - 3*ULN	3*ULN - 5*ULN	5*ULN - 20*ULN	20*ULN	1.5*BLN - 3.0*..	3.0*BLN - 5.0*..	5*BLN - 20*BLN	20*BLN
6	BIOCHEMISTRY	Bilirubin						ULN - 1.5*ULN	1.5*ULN - 3*UL..	3*ULN - 10*ULN	10*ULN	BLN - 1.5*BLN	1.5*BLN - 3*BL..	3*BLN - 10*BLN	10*BLN
7	BIOCHEMISTRY	Calcium, Ionized	mmol/L	LLN - 1	1.0 - 0.9	0.9 - 0.8	0.8	ULN - 1.5	1.5 - 1.6	1.6 - 1.8	1.8				
8	BIOCHEMISTRY	Calcium Corrected	mmol/L	LLN - 2	2 - 1.75	1.75 - 1.5	1.5	ULN - 2.9	2.9 - 3.1	3.1 - 3.4	3.4				
9	BIOCHEMISTRY	Calcium	mmol/L	LLN - 2	2 - 1.75	1.75 - 1.5	1.5	ULN - 2.9	2.9 - 3.1	3.1 - 3.4	3.4				
10	BIOCHEMISTRY	Creatine Kinase						ULN - 2.5*ULN	2.5*ULN - 5*UL..	5*ULN - 10*ULN	10*ULN				
11	BIOCHEMISTRY	Creatinine						ULN - 1.5*ULN	1.5*ULN - 3*UL..	3*ULN - 6*ULN	6*ULN		1.5*BLN- 3*BLN	3*BLN	
12	BIOCHEMISTRY	Gamma Glutamyl Transferase						ULN - 2.5*ULN	2.5*ULN - 5*UL..	5*ULN - 20*ULN	20*ULN	2.0*BLN - 2.5*..	2.5*BLN - 5*BL..	5*BLN - 20*BLN	20*BLN

LBCAT	LBTESTCD	DIRECTION	LBSTRESU	CRITERIA	LBTOXGR	LBTOX
BIOCHEMISTRY	CREAT		UMOL/L	lbstresn<=1*lbstnrhi	0	
BIOCHEMISTRY	CREAT	HIGH	UMOL/L	1*lbstnrhi<lbstresn<=round(1.5*lbstnrhi,1e-12)	1	Creatinine increased
BIOCHEMISTRY	CREAT	HIGH	UMOL/L	round(1.5*lbstnrhi,1e-12)<lbstresn<=round(3*lbstnrhi,1e-12)	2	Creatinine increased
BIOCHEMISTRY	CREAT	HIGH	UMOL/L	round(1.5*base,1e-12)<lbstresn<=round(3*base,1e-12) and lbstresn>lbstnrhi	2	Creatinine increased
BIOCHEMISTRY	CREAT	HIGH	UMOL/L	round(3*lbstnrhi,1e-12)<lbstresn<=round(6*lbstnrhi,1e-12)	3	Creatinine increased
BIOCHEMISTRY	CREAT	HIGH	UMOL/L	round(3*base,1e-12)<lbstresn and lbstresn>lbstnrhi	3	Creatinine increased
BIOCHEMISTRY	CREAT	HIGH	UMOL/L	round(6*lbstnrhi,1e-12)<lbstresn	4	Creatinine increased

BeOne SDTM Status

BeOne SDTM Status

Form OID	Form Name	Field Order	Field OID	Field Label	Field Option Order	Field Option	SDTM Domain	SDTM Variable	SDTM Variable Value	SDTMRecordID	Algorithm
VS_HORIZONTAL	Vital Signs HorizontalGeneric	1.1	VSPERF	Were vital signs performed?	1	Yes					NotSubmitted
VS_HORIZONTAL	Vital Signs HorizontalGeneric	1.1	VSPERF	Were vital signs performed?	2	No	VS	VSTESTCD	VSALL	VSPERF	ifAssigned
VS_HORIZONTAL	Vital Signs HorizontalGeneric	1.1	VSPERF	Were vital signs performed?	2	No	VS	VSTEST	Vital Signs	VSPERF	ifAssigned
VS_HORIZONTAL	Vital Signs HorizontalGeneric	1.1	VSPERF	Were vital signs performed?	2	No	VS	VSSTAT	NOT DONE	VSPERF	ifAssigned
VS_HORIZONTAL	Vital Signs HorizontalGeneric	1.2	VSDAT	Date (DD-MMM-YYYY)			VS	VSDTC	YYYY-MM-DD		DT2DTC
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.1	SYSBP_VSORRES	Systolic Blood Pressure			VS	VSTESTCD	SYSBP	SYSBP_VSORRES	Assigned
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.1	SYSBP_VSORRES	Systolic Blood Pressure			VS	VSTEST	Systolic Blood Pressure	SYSBP_VSORRES	Assigned
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.1	SYSBP_VSORRES	Systolic Blood Pressure			VS	VSORRES		SYSBP_VSORRES	Copy
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.2	SYSBP_VSORRESU	Systolic Blood Pressure Unit			VS	VSORRESU		SYSBP_VSORRES	Copy
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.3	DIABP_VSORRES	Diastolic Blood Pressure			VS	VSTESTCD	DIABP	DIABP_VSORRES	Assigned
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.3	DIABP_VSORRES	Diastolic Blood Pressure			VS	VSTEST	Diastolic Blood Pressure	DIABP_VSORRES	Assigned
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.3	DIABP_VSORRES	Diastolic Blood Pressure			VS	VSORRES		DIABP_VSORRES	Copy
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.4	DIABP_VSORRESU	Diastolic Blood Pressure Unit			VS	VSORRESU		DIABP_VSORRES	Copy
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.5	BP_VSPOS	Blood Pressure Position	1	Prone	VS	VSPOS	PRONE	SYSBP_VSORRESI DIABP_VSORRES	CopyUpcase
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.5	BP_VSPOS	Blood Pressure Position	2	Semi-Recumbent	VS	VSPOS	SEMI-RECUMBENT	SYSBP_VSORRESI DIABP_VSORRES	CopyUpcase
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.5	BP_VSPOS	Blood Pressure Position	3	Sitting	VS	VSPOS	SITTING	SYSBP_VSORRESI DIABP_VSORRES	CopyUpcase
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.5	BP_VSPOS	Blood Pressure Position	4	Standing	VS	VSPOS	STANDING	SYSBP_VSORRESI DIABP_VSORRES	CopyUpcase
VS_HORIZONTAL	Vital Signs HorizontalGeneric	2.5	BP_VSPOS	Blood Pressure Position	5	Supine	VS	VSPOS	SUPINE	SYSBP_VSORRESI DIABP_VSORRES	CopyUpcase

BeOne Real-time SDTM Status

- **50+** ongoing studies in SDTM pipeline
- **44** ongoing studies in SDTM daily auto-run
- ~50% Studies' SDTM Auto-run Success every day
  - No Log Issues
  - Data are compared in double programing
- SDTM Team: ~6.5 FTEs in total including SDTM define package
  - 2-2.5 FTEs at Global Level
  - 4-4.5 FTEs at Study Level
- **SDTM Work Model:** Just like ADaM/TFL programmers, SDTM programmers are also team members led by the Study Lead Programmer.
  - Triple programming is not required.
  - A risk-based oversee is necessary.

Challenges of Conducting Data Cleaning Based on SDTM

Challenges of Conducting Data Cleaning Based on SDTM

- Tracing from SDTM to Raw Data: How to establish a complete loop from SDTM back to raw data.
- DM Understanding of SDTM Domains: Ensuring DM can quickly trace corresponding raw records from SDTM results for query checks.
- Validation of SDTM Results: How to ensure SDTM results are validated enough to fully replace current raw data checks. If not sufficiently validated, SDTM-based reviews remain supplementary.
- Error Impact in SDTM: Evaluating the acceptance of potential errors in SDTM affecting issue checks and strategies to mitigate this risk.
- Delay in SDTM Refresh: Addressing the delay in updating SDTM with raw data.

Future Outlook

Future Outlook

Incremental SDTM Generation: Generate SDTM based on incremental data rather than the entire dataset to accelerate output. This approach also reduces the time required for issue verification based on SDTM.

AI Empowerment: Utilize AI to automatically generate query reports based on standard libraries and write queries automatically using historical data.

Automated Query Entry: Simulate web operations to automatically fill queries into the EDC system, reducing manual work.

Risk-Based Query Review: Analyze trends in queries from previous projects to identify sites and queries that occur infrequently, reducing checks on such data to improve efficiency.

Thank You!