PharmaSUG China 2025 - Paper AA-112

Dynamic Generation of Study-adaptive SDTM Derived Variables and Corresponding Specifications in SDTM Automation

Qiuan Hu, Chunpeng Zhao, Liming Xie, BeOne Medicines Ltd.

ABSTRACT

In the clinical programming industry, the traditional generation of SDTM datasets has relied heavily on SAS programmers to manually develop annotated Case Report Forms (aCRF), SDTM programming specifications, and domain-specific SAS programs. While existing automation solutions have improved efficiency in certain discrete steps such as automated aCRF generation and CRF data to SDTM mapping, significant challenges persist in dynamically generating study-adaptive SDTM derived variables (e.g., xxSEQ, xxLOBXFL, EPOCH) and their corresponding specifications.

In BeiGene, we have developed an integrated automation suite for standardized SDTM deliverables that comprehensively addresses aCRF generation, Trial Design Model (TDM) implementation, dynamically CRF SDTM minus data processing, and dynamic study-adaptive SDTM derived data creation to realize the real-time SDTM. This paper details our innovative approach to the dynamic generation of study-adaptive SDTM derived variables and corresponding specifications, with the ultimate goal of achieving the full automation of derived data within the SDTM automation transformation process.

INTRODUCTION

According to the Study Data Tabulation Model Implementation Guide (SDTM-IG), derived data plays a vital role in SDTM for each domain. A critical challenge lies in developing an automated approach to define the scope of derived variables for each domain, while eliminating dependency on data-driven methods for real-time generation.

This paper introduces a metadata-driven transformation architecture that enables study-adaptive SDTM derived variables and corresponding specifications through systematic integration of information from the Trial Design Model (TDM), corresponding Architect Loader Specification (ALS) document, Case Report Form (CRF) and external Data Transfer Specification (DTS) files. Programs for derived data also can be efficiently generated based on these corresponding specifications, thereby achieving the automated transformation from raw data to SDTM independent of actual raw data availability.

REAL-TIME SDTM TRANSFORMATION AND DATA CLASSIFICATION

The rationality of real-time SDTM is that the SDTM transformation program is not data driven. To enable real-time SDTM, both SDTM transformation specifications and SDTM transformation code must be prepared prior to the First Patient First Visit (FPFV). Once raw data is available, it can be immediately transformed into SDTM.

In real-time SDTM, data in each SDTM record can be conceptually fragmented into four core components: Trial Design Model (TDM) data, CRF-captured data (also known as CRF minus data), external data (also known as eDT minus data or DTS data) and derived data.

Trial Design Model (TDM) data is captured from corresponding Architect Loader Specification (ALS) files and study protocol to create the clinical trial design domains (e.g., *TA TE TV TS*), which provide a brief, clear description of the plan and design of the study.

CRF minus datasets are the mini-SDTM version compared with final SDTM datasets, which mainly include all the variables from aCRF, the MedDRA/WHODrug coding variables, and some EDC system default variables. The process of creating a CRF minus dataset involves CRF data information remapping or variable renaming in SDTM-like format, ISO datetime formatting or standard class code variables, test or unit conversion, and so on. Figure 1 is an example, CRF data are firstly directly transformed and renamed to SDTM variables *LBNAM* & *LBFAST* & *LBORNRLO* & *LBORNRHI* & *LBCLSIG*; test and result data are mapped to SDTM test synonym qualifier and result qualifier variables *LBTESTCD & LBTEST & LBORRES & LBORRESU*; then SDTM ISO 8601 datetime or interval timing variable *LBDTC*; finally

standard test or unit conversion variables *LBSTRESC & LBSTRESN & LBSTRESU & LBSTNRLO & LBSTNRHI* and corresponding reference range indicator variables *LBNRIND* are created based on CRF raw result and range variables. The CRF minus data transformation specification (also known as crf2sdtm spec) is created centrally by the SDTM Annotation Centralization Group at the CRF design stage. Once the first CRF data record is available, it is immediately transformed into the CRF minus dataset.

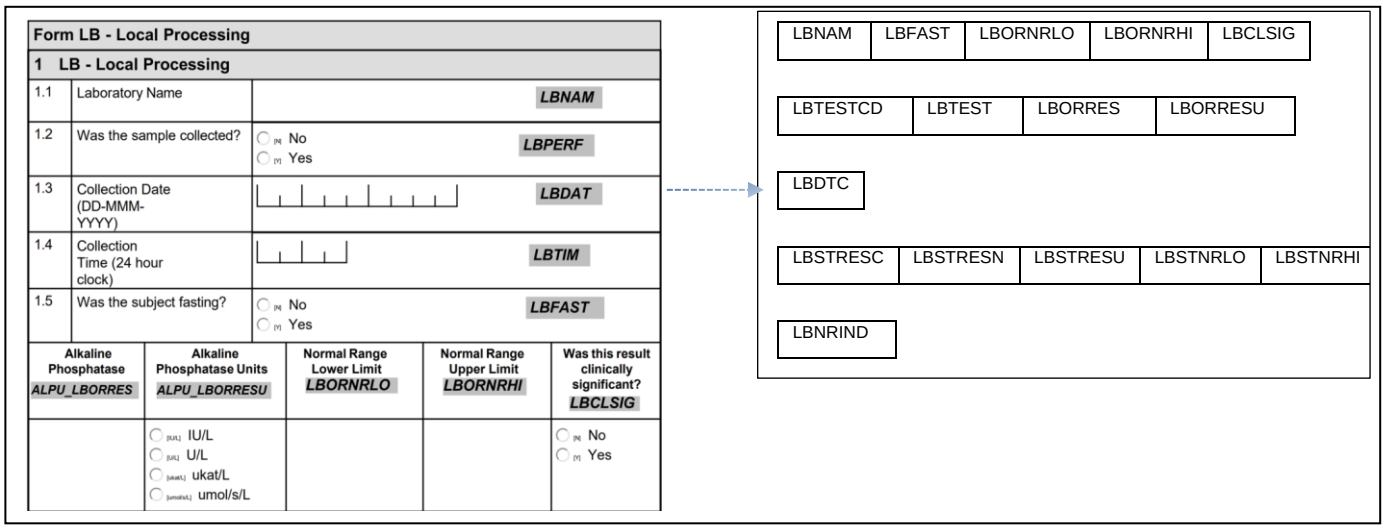

Figure 1. Example of LB minus dataset based on Form LB - Local Processing on the CDISC website

The eDT minus data is sourced from external vendors. For all studies, Data Transfer Specification (DTS) templates are standardized and reviewed by the SDTM DTS Review Centralization Group, when available. After the DTS is reviewed, the SDTM transformation specification (also known as edt2sdtm spec) for eDT minus data is centrally created by the SDTM Annotation Centralization Group. Based on the edt2sdtm spec, study SDTM owner is responsible for generating the eDT minus dataset.

Derived data refers to standardized information that is neither part of CRF minus data nor eDT minus data. These commonly include variables such as *EPOCH, LBTOX, VISIT, VISITNUM, xxDY, xxBLFL, xxSEQ*, among others. The scope of derived variables is dynamically determined based on specific domain requirements outlined in the SDTM Implementation Guide (SDTM-IG), guidance from the FDA and CDISC standards, as well as crf2sdtm spec and edt2sdtm spec.

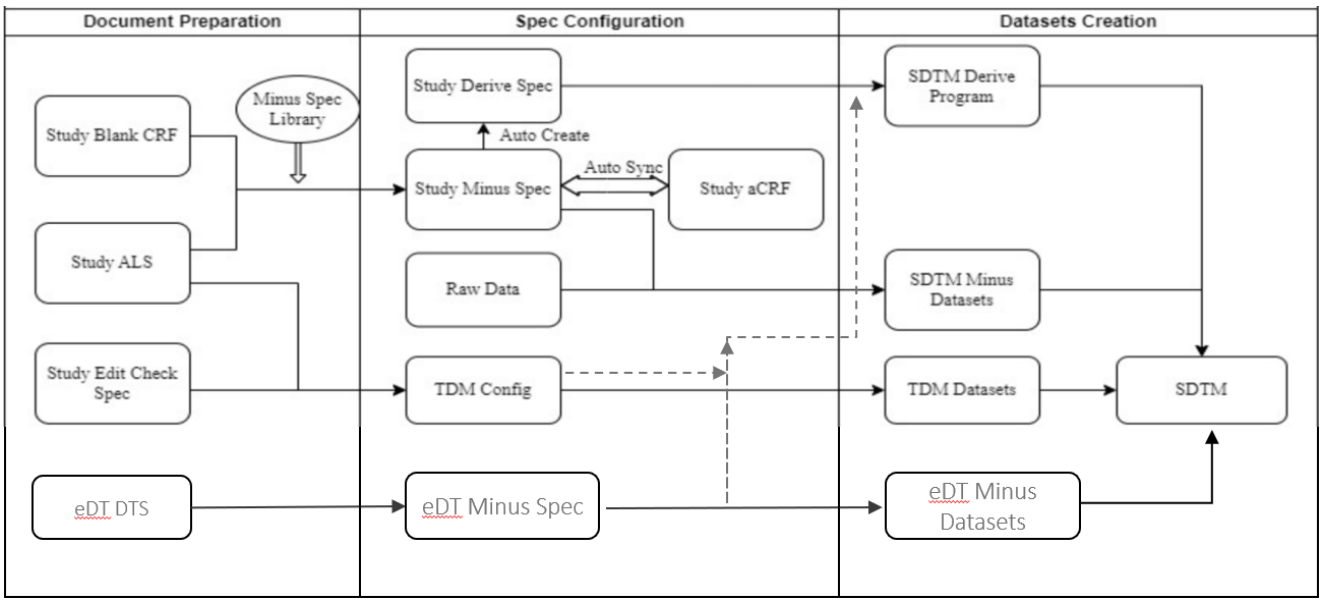

Figure 2. Flowchart of SDTM Transformation

SDTM DERIVED DATA DYNAMIC GENERATION

SDTM derived variables and specifications dynamic generation

Once CRF minus, eDT minus and TDM data along with corresponding specifications are prepared, the final step is to generate SDTM datasets for each domain. This involves adding SDTM derived variables to the minus datasets, generating SDTM programs, and transforming them into the final SDTM datasets. This is a dynamic process, as it may involve adding or removing domains (e.g., *SE, SV*) or derived variables to align with the specific characteristics of the study data and clinical trial design to customize the study-adaptive derived specifications (also known as derived spec), which contain the programming algorithm for generating SDTM derived variables based on ALS, TDM, SDTM minus, eDT minus dataset and a global fixed template derived spec and a set of standardized global macros.

The derived spec outlines the detailed list of derived variables, macros used for each variable, and the macro algorithm descriptions for each domain, following a predefined dynamically updated order. It is usually only the required and expected variables defined in SDTM-IG are added as long as they are not already present in SDTM minus except for EX and DM domain since they are respectively derived from EC and DC domain. Each derived variable or a group of related derived variables within a domain is derived together within a single program or macro by a shared variable order number to ensure consistent processing and traceability.

Figure 3 illustrates an example of a derived variable and the corresponding derived spec for the LB domain. Since LB is a findings domain, timing variables such as *VISIT* & *VISITNUM*, as well as baseline variables like *LBBLFL* & *LBLOBXFL* are typically derived. Additionally, The EPOCH variable and study day variable (LBDY) are also triggered to derive when the CRF includes Specimen Collection Date/Time information. Toxicity Grade variables (*LBTOX* & *LBTOXGR*) are special qualifier variables for the LB domain. The variable *LBTPTNUM* is derived based on *LBTPT* collected from CRF. *LBSEQ* is a required Identifier variable as defined in the SDTM-IG. *VISIT* & *VISITNUM, LBTOX* & *LBTOXGR* sharing the same VariableOrder number within one program or macro since they are a group of related variables.

Domain Orde	Domain	Variable Orde	Variable	VariableLabel	AlgorithmLabel	QCAlgorithmLabel	AlgorithmDescription
190	LB	10	VISIT	Visit Name	CM_POST CM_PST DM IE MH_MCL PR_PC PR_POS PR_POSR	CM_BPT CM_POST CM_PST DM IE MH_MCL PR_PC PR_POS PR_POSR	$visit.); For unscheduled visit, merge SV on INSTANCEREPEATNUMBER=SV.SVSPID to get
190	LB	10	VISITNUM	Visit Number	CM_POST CM_PST DM IE MH_MCL PR_PC PR_POS PR_POSR	CM_BPT CM_POST CM_PST DM IE MH_MCL PR_PC PR_POS PR_POSR	Merge with SV by USUBJID and VISIT to get VISITNUM
190	LB	20			lb-edt	lb-edt-qc	exteral eDT data mapping and combine process for LB domain
190	LB	30	EPOCH	Epoch	%m_sdtm_epoch	%m_sdtmqc_epoch	Derived from SE domain based on LBDTC, SE.SESTDTC and SE.SEENDTC.
190	LB	40	LBDY	Study Day of Specimen Collection	%m_sdtm_studyday	%m_sdtmqc_studyday	LBDTC-RFSTDTC+(LBDTC>=RFSTDTC)
190	LB	50	LBENDY	Study Day of End of Observation	%m_sdtm_studyday	%m_sdtmqc_studyday	LBENDTC-RFSTDTC+(LBENDTC>=RFSTDTC)
190	LB	60	LBTPTNUM	Planned Time Point Number	%m_sdtm_map(source=LBTPT,map="PRE-DOSE"#1|"4-6 HOURS POST-DOSE"#2)	%m_sdtmqc_map(source=LBTPT,map="PRE-DOSE"#1|"4-6 HOURS POST-DOSE"#2)	to be re-numbered to set all the timepoints in choronological order)
190	LB	70	LBBLFL	Baseline Flag	%m_sdtm_blfl(rfsdtvar=RFSTDTC)	%m_sdtmqc_blfl(rfsdtvar=RFSTDTC)	LBBLFL=Y for last non missing record on or before the reference start date (RFSTDTC).
190	LB	80	LBLOBXFL	Last Observation Before Exposure Flag	%m_sdtm_blfl(rfsdtvar=RFXSTDTC)	%m_sdtmqc_blfl(rfsdtvar=RFXSTDTC)	Flag "Y" if it's the last non-missing value prior to RFXSTDTC.
190	LB	90	LBTOX	Toxicity	%m_sdtm_lbtox	%m_sdtmqc_lbtox	is imputed by add or substract 0.0000000001 respectively before calculating the toxicity.
190	LB	90	LBTOXGR	Standard Toxicity Grade	%m_sdtm_lbtox	%m_sdtmqc_lbtox	is imputed by add or substract 0.0000000001 respectively before calculating the toxicity.
190	LB	999	LBSEQ	Sequence Number	VISITNUM LBDTC LBCAT LBSCAT LBTESTCD LBGRPID LBSPEC LBTPTNUM	VISITNUM LBDTC LBCAT LBSCAT LBTESTCD LBGRPID LBSPEC LBTPTNUM	2. Assigned as 1 for the first record of the subject 3. Incremented by 1 for each subsequent record

Figure 3. Example of derived variables and the corresponding specifications for LB domain

The algorithms embedded in the macros are pre-defined, and their parameters can be automatically populated and refreshed using the information from SDTM minus spec, eDT minus spec, TDM data, and ALS file. For example, in Rave EDC system, a visit is identified by the folder name and the values of folder name is listed in the study's ALS document, so ALS can helps determine which folder names and CRF forms (EDC source dataset) are supported to derive the SV domain. It also identifies which domains should trigger the derivation of VISIT-related variables, based on domain-specific characteristics. This information is dynamically captured and automatically filled into the macro parameters, enabling efficient and consistent derivation of programming algorithms of SDTM variables across studies.

ALS file:

Matrix: MASTERDAS HBOARD	SCR	C1D1	C1D22	C2RW1D1	C2RW2D1
CE BSYM	X	X	X	X	
LB CHEM	X	X	X	X	
LB COAG	X				
LB HEMA	X	X	X		
EC OR5				X	X
AE MIX					
CM POST					
PR POS					

OID	FolderName
AE	Adverse Events
C1D1	Cycle 1 Day 1
C1D22	Cycle 1 Day 22
C2RW1D1	Cycle 2/Ramp-up Week 1 Day 1
C2RW2D1	Cycle 2/Ramp-up Week 2 Day 1
SCR	Screening

Macro parameters for VIST variable derived in SV and LB domain:

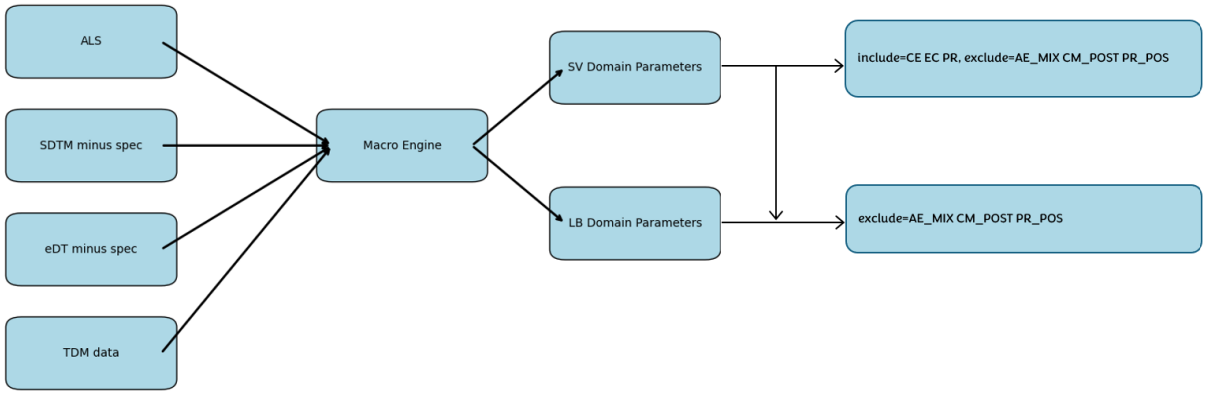

Figure 4. Example of CRF form and foldername in ALS file

Specification meta-programming and study-adaptive refresh

After the study derived specification and the programming algorithms for all related SDTM derived variables in each domain are created, users can generate all SDTM programs and final SDTM datasets by running a script or macro. This script or macro consolidates all sub-programs, macros and assignments into a single program to batch run all domains based on the study derived specification. Additionally, if users only want to generate a specific domain's derived data program, they can specify the domain list when running the macro or script to re-generate the corresponding SDTM program and final SDTM dataset for development or validation purposes. Furthermore, the process can automatically split the dataset into the main SDTM and SUPP domains.

The programming approach described above for each domain is a form of meta-programming. Meta-programming takes SAS code as data, and the direct outputs of meta-programming are not SDTM data sets, but rather metadata and SAS programs. The key to meta-programming is metadata. Metadata is a universal data model, which is applicable to all SDTM derived variables and their corresponding SDTM domains. Importantly, the metadata is machine-readable, which has two implications. First, metadata can be recycled and reused through code. Second, executable SAS programs (e.g., lb.sas in the example below) can be automatically generated from metadata, rather than manual coding. This is possible because all the information required to create these programs is embedded in the machine-readable study-adaptive derived specification.

In SDTM derived data transformation, the derived specification serves as metadata, representing programming algorithms. SAS is used solely to interpret this metadata and generate executable SAS programs, functioning as a compilation process to produce SDTM datasets. The metadata, programming algorithms, and transformation rules can be dynamically refreshed in real-time based on study-adaptive CRF forms, fields, and study design. Moreover, meta-programming enables a coding-free approach to SDTM derived data transformation and lays the foundation for SDTM derived data full automation.

```
/*Define global macro variables dsin and dsout, representing input dataset name and output dataset name
respectively in each step.*/
%global dsin dsout;

/*Derive VISIT/VISITNUM from the macro.*/
%m_sdtm_visit(exclude=AE_MIX CM_POST PR_POS,dsin=sdtm.lb_,dsout=_tmplb1)
/*VISIT/VISITNUM derivation ends.*/

/*Pre-process for the domain lb external data.*/
%let dsin=_tmplb1;
%let dsout=_tmplb2;
data &dsout.;
   set sdtm.lb_edt;
run;
/*Pre-process for the domain lb external data ends.*/

/*Combine edc data and external data for the domain lb.*/
%m_sdtm_combine(dsin1=_tmplb1,dsin2=_tmplb2,dsout=_tmplb3)
/*Combination of edc data and external data for the domain lb ends.*/

/*Derive EPOCH from the macro.*/
%m_sdtm_epoch(target=EPOCH,dsin=_tmplb3,dsout=_tmplb4)
/*EPOCH derivation ends.*/

/*Derive LBDY from the macro.*/
%m_sdtm_studyday(target=LBDY,dsin=_tmplb4,dsout=_tmplb5)
/*LBDY derivation ends.*/

/*Derive LBENDY from the macro.*/
%m_sdtm_studyday(target=LBENDY,dsin=_tmplb5,dsout=_tmplb6)
/*LBENDY derivation ends.*/

/*Derive LBTPTNUM from the macro.*/
%m_sdtm_map(source=LBTPT,map="PRE-DOSE"#1|"4 HOURS POST-DOSE"#2,target=LBTPTNUM,
dsin=_tmplb6,dsout=_tmplb7)
/*LBTPTNUM derivation ends.*/

/*Derive LBBLFL from the macro.*/
%m_sdtm_blfl(rfsdtvar=RFSTDTC,target=LBBLFL,dsin=_tmplb7,dsout=_tmplb8)
/*LBBLFL derivation ends.*/

/*Derive LBLOBXFL from the macro.*/
%m_sdtm_blfl(rfsdtvar=RFXSTDTC,target=LBLOBXFL,dsin=_tmplb8,dsout=_tmplb9)
/*LBLOBXFL derivation ends.*/

/*Derive LBTOX/LBTOXGR from the macro.*/
%m_sdtm_lbtox(dsin=_tmplb9,dsout=_tmplb10)
/*LBTOX/LBTOXGR derivation ends.*/

/*Derive LBSEQ from the macro.*/
%m_sdtm_seq(keyvar=STUDYID USUBJID VISITNUM LBDTC LBCAT LBSCAT LBTESTCD LBGRPID LBSPEC
LBTPTNUM LBORRES,target=LBSEQ,dsin=_tmplb10,dsout=_tmplb11)
/*LBSEQ derivation ends.*/

/*Split it into main domain and SUPPQUAL and store it permanently*/
%m_sdtm_create(ig_version=3.3,dsin=_tmplb11,dsout=lb,outlib=sdtm,metadata=location/ xxx_derived_spec.xlsx)
```

Figure 5. Example of Generate Executable SAS Program Based on derived Specification Metadata

THE RATIONALITY OF SDTM DERIVED DATA FULL AUTOMATION

The derivation of SDTM derived domains and variables is meta-programming process based on SDTM derived specifications. The specifications are generated based on TDM, CRF SDTM minus data and eDT SDTM minus data. These sources collectively determine which SDTM derived domains and variables should be created. The roles and key derivation rules for SDTM derived domains and variables are defined in the SDTM-IG and remain consistent across all studies. They are not influenced by the study protocol or the statistical analysis plan (SAP). All study specific setting can be done in TDM, CRF and DTS, which correspond respectively to TDM data, CRF SDTM minus data and eDT SDTM minus data. In essence, the derivation of SDTM derived data is a meta-programming process that integrates TDM data, CRF SDTM minus data and eDT SDTM minus data. SDTM derived specifications and SDTM derived Code/macros are taken as one same thing and are considered two representations of the same logic. The differences are that derived specifications are intended for human review, while code/macros for machine read and execution. When CR migrations, study design change or updates to eDT data occur during the study cycle, the transformation of SDTM derived data can be automatically refreshed through the automatically updated derived specifications and macro/code, which enables the full automation of SDTM derived data generation process.

CONCLUSION

The ultimate goal of dynamic generation of study-adaptive SDTM derived variables and specifications is to achieve full automation of SDTM derived data, which is a key component of the real-time SDTM derived data full automation. In this process, derived specifications serve as essential programming metadata for the SDTM derived data transformation. Once the metadata model is stabilized, SDTM derived data transformation becomes coding-free and full automation, as executable SAS programs can be directly automatically generated from metadata rather than manual coding. Study-adaptive SDTM transformation specifications corresponding transformation code are prepared prior to First Patient First Visit (FPFV). As soon as the first patient data record is available, it is immediately transformed into SDTM format. When CR migrations, CRF annotations, or study design changes occur during the study cycle, the SDTM derived data transformation code can be still automatically refreshed by the automatically updated specifications metadata, ensuring the real-time SDTM delivery.

REFERENCES

Nana Yang, Chunpeng Zhao. “The Imperative Role of Data Cut-Off (DCO) in Clinical Trials: Ensuring Integrity and Precision in Oncology”. PharmaSUG China 2024, DS10092. Available at https://www.lexjansen.com/pharmasug-cn/2024/DS/Pharmasug-China-2024-DS10092_Final_Paper.pdf.

eCRF Portal. https://www.cdisc.org/kb/ecrf/laboratory-test-results-local-processing.

Chunpeng Zhao, Mijun Hu, Jie Liu, Jiapeng He, Liming Xie, Aide Zhou, Bingting Tao. “Real-time SDTM”. PharmaSUG China 2022, AT105. Available at https://www.lexjansen.com/pharmasug-cn/2022/AT/Pharmasug-China-2022-AT105.pdf.

Mingyue Zhu, Wei Wang, Mingfeng Zhou. “An Automatic SDTM Platform: Create Your SDTM Since Case Report Form is Ready!”. PharmaSUG China 2024, AA10053. Available at https://www.lexjansen.com/pharmasug-cn/2024/AA/Pharmasug-China-2024-AA10053_Final_Paper.pdf.

Qian SU and Michael WANG, Sanofi. “Metadata-Driven Programming Using R: A Case Study in BDS”. PharmaSUG China 2024, CC10011. Available at https://www.lexjansen.com/pharmasug-cn/2024/CC/Pharmasug-China-2024-CC10011_Final_Paper.pdf.

Haiqiang Luo, Kaiying Li, Xiaoyi Niu, Xiaoxia Zhou. “SDTM ARTs - a CRF-Driven SDTM Automation Tool for Generating aCRF, SDTM Mapping Specification and Programs”. PharmaSUG China 2021, AD67. Available at https://www.lexjansen.com/pharmasug-cn/2021/AD/Pharmasug-China-2021-AD067.pdf

Chunpeng Zhao, Mijun Hu, Jiapeng He, Aide Zhou. “SDTM Annotated CRF Digitalization”. PharmaSUG China 2021, DS036. Available at https://www.lexjansen.com/pharmasug-cn/2021/DS/Pharmasug-China-2021-DS036.pdf

ACKNOWLEDGMENTS

Thanks to our managers (Leo Li, Cindy Song and Tingting Zeng) for their great support on the conference and programming colleagues for their help on preparing the paper.

RECOMMENDED READING

- *CDISC SDTMIG v3.3, CDISC SDTMIG v3.4, SDTMIG-PGx v1.0*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Qiuan Hu
BeOne Medicines Ltd.
qiuan.hu@beigene.com

Chunpeng Zhao
BeOne Medicines Ltd.
chunpeng.zhao@beigene.com

Liming Xie
BeOne Medicines Ltd.
liming.xie@beigene.com