How does SDTM Programming Work?

HyunSoo Lee, Associate Partner, JNPMEDI

Meet the Speaker

HyunSoo Lee

Title: Associate Partner

Organization: JNPMEDI

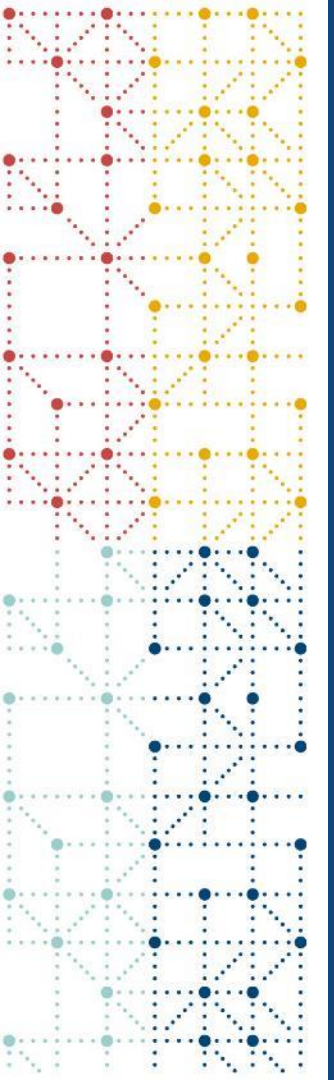
HyunSoo Lee has more than 20 years professional experience in Data Management and Biostatistics having had the opportunity to work for several pharmaceutical companies and CROs.

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. Understanding the Basic Concepts
2. What is SDTM Programming?
3. How do you convert raw data to SDTM format?
4. Maven Converter



1. Understanding the Basic Concepts

What is CDISC?
What is SDTM in CDISC?

CDISC = Power of Standardization

Clinical **D**ata **I**nterchange **S**tandard **C**onsortium (**CDISC**)

- Global **S**tandards **D**evelopment **O**rganization (SDO)
- Founded in 1997 (all volunteers)
- Incorporated in 2000 as a non-profit organization

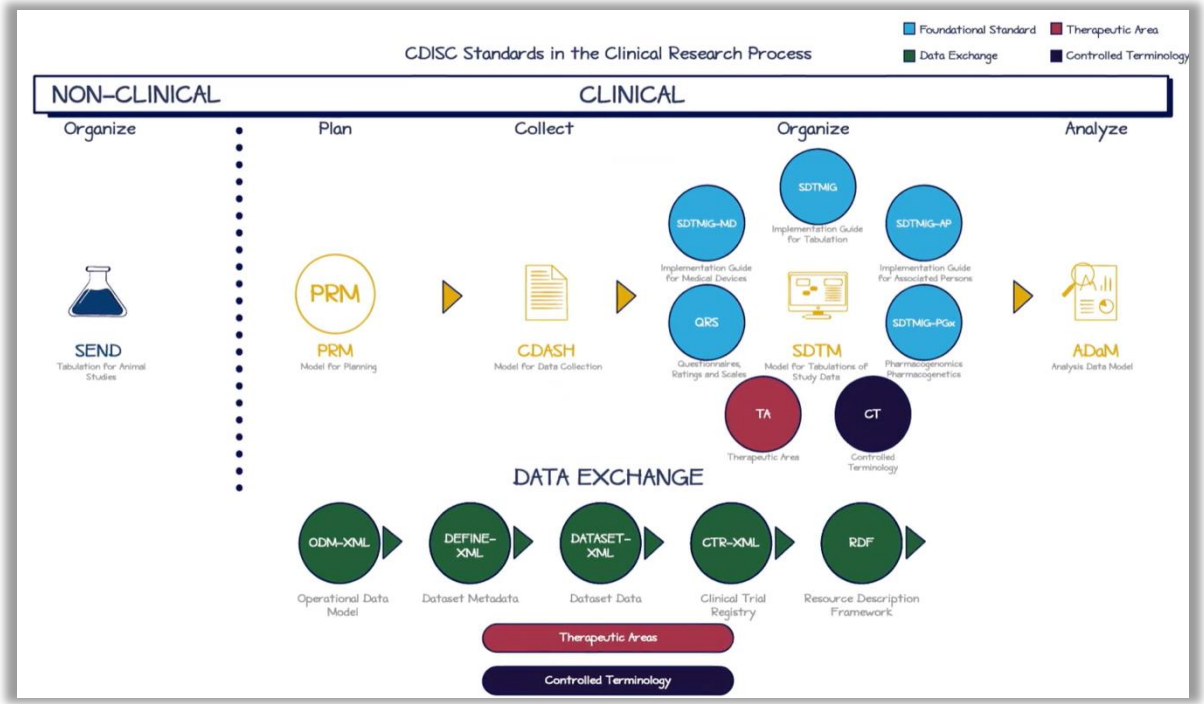
Study Data Tabulation Model (**SDTM**) defines a **standard structure** for human clinical trial (study) data tabulations and for non-clinical study data tabulations, that are to be submitted as part of a product application, to a regulatory authority such as the United States Food and Drug Administration (FDA).

ICONS from FLATICON

CDISC Standards in the Research Process

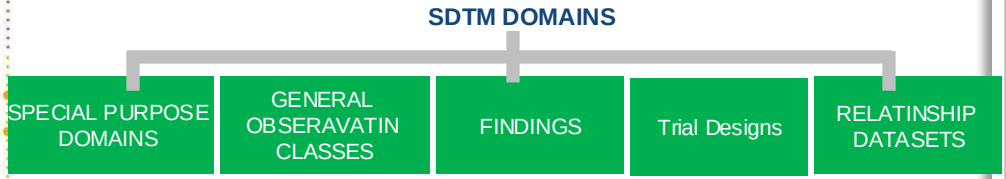
Capture CDISC Training Materials

CDISC Standards



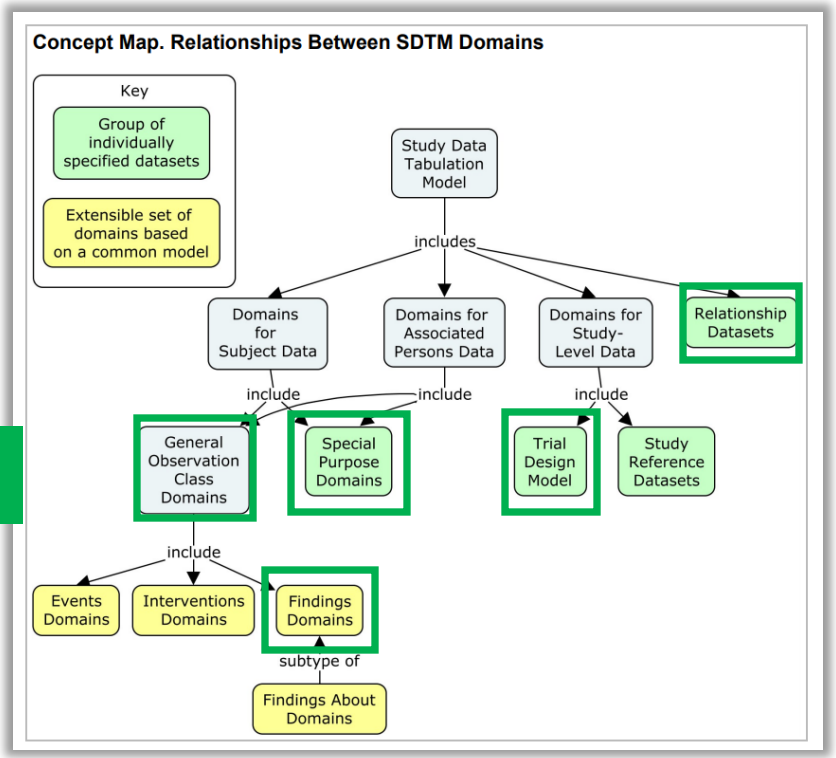
Capture CDISC Training Materials

Organization of the SDTM



Capture & Text from CDISC

CDISC Study Data Tabulation Model (2.0 Final)



Domain vs. Dataset: What's the Difference?

The terms "domain" and "dataset" are commonly used in CDISC's nomenclature and found frequently in the **Study Data Tabulation Model (SDTM)**. For example, the **SDTM v1.8** includes 134 instances of "domain" and says "A collection of observations on a particular topic is considered a domain." The Model includes 78 instances of dataset and certain structures in the model are called "datasets" rather than "domains." Is there a difference between a domain and a dataset?

The **CDISC Glossary** defines these terms as follows:

- Domain: A collection of logically related observations with a common, specific topic that are normally collected for all subjects in a clinical investigation. NOTE: The logic of the relationship may pertain to the scientific subject matter of the data or to its role in the trial. Example domains include laboratory test results (LB), adverse events (AE), concomitant medications (CM). [After SDTM Implementation Guide version 3.2, **CDISC.org**] See also general observation class.
- Dataset: A collection of structured data in a single file. [CDISC, ODM, and SDS] Compare to analysis dataset, tabulation dataset

A domain is a collection of observations on a particular topic.

A dataset is a collection of structed data in a single file.

Capture & Text from CDISC

FDA, NMPA, & PMDA Regulatory Requirements

Standard	FDA - US	PMDA - Japan	NMPA
Controlled Terminology	★		
SEND	★		
SDTM	★	★	★
ADaM	★	★	★
Define-XML	★	★	
Analysis Results Metadata (ARM)		★	

Capture CDISC Training Materials

FDA, NMPA, & PMDA Regulatory Requirements

https://www.cdisc.org/video/regulatory-requirements

Agency	Clinical CDISC	Nonclinical CDISC
EMA	Not referenced	Not referenced
FDA	SDTM – CDER[1)] & CBER[2)] – Required – 12/17/2016* ADAM – CDER & CBER – Required – 03/15/2019*	CBER – Required – 03/15/2023 CDER – Required – 12/17/2016*
NMPA	Preferred – September 2019	No Requirement
PMDA	SDTM & ADaM – Required – 04/01/2020	No Requirement

1) Center for Drug Evaluation and Research
2) Center for Biologics Evaluation and Research
*Requirements are dependent on type of submission, IND, ANDA, NDA, BLA, etc.

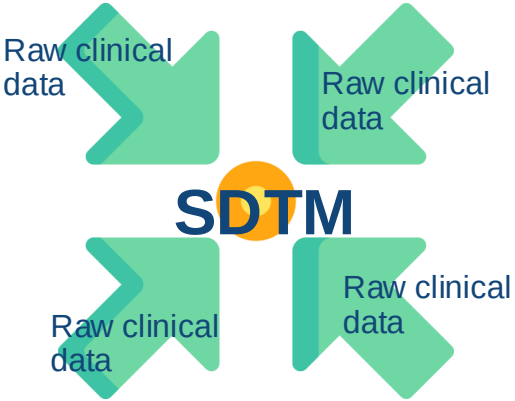
What is SDTM Programming?

SDTM Programming

SDTM Programming is the process of transforming raw clinical trial data into a standardized format that is used for regulatory submission and data analysis.

SDTM Programming

SDTM Programming is the backbone of data organization in clinical trials.



- ✓ 표준화형식으로 변환하여 FDA와 같은 규제기관에 더 용이하게 제출.
- ✓ SDTM Structure는 다양한 출처와 연구 데이터를 일관되게 해석하고 비교 할 수 있는 방식으로 설정
- ✓ 임상시험 결과의 무결성과 신뢰성을 유지하는 것이 필수

ICONS from FLATICON

SDTM Mapping?

SDTM Programming?

ICONS from FLATICON

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

SDTM Programming

Data Collection ✓ Typically originates from EDC systems used in clinical trials.

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

SDTM Mapping Process

Analyzing the Raw Data

Define the SDTM Domains

Mapping Raw Data to SDTM Domains

Handling Controlled Terminology

Validation

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

SDTM Mapping Process

Analyzing the Raw Data

- ✓ Clinical trial data, Labs, Medical Devices, 환자가 보고한 자료.
- ✓ Raw Data를 명확히 이해하는 것이 목적임.

Define the SDTM Domains

Mapping Raw Data to SDTM Domains

Handling Controlled Terminology

Validation

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

SDTM Mapping Process

Analyzing the Raw Data

Define the SDTM Domains

- ✓ DM, AE, LB, MH 등을 정의함.
- ✓ Choose the relevant SDTM domains based on the type study.

Mapping Raw Data to SDTM Domains

Handling Controlled Terminology

Validation

ICONS from FLATICON

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

SDTM Mapping Process

Analyzing the Raw Data

Define the SDTM Domains

Mapping Raw Data to SDTM Domains

✓ How do you map it with any tool??

Handling Controlled Terminology

Validation

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

SDTM Mapping Process

Analyzing the Raw Data

Define the SDTM Domains

Mapping Raw Data to SDTM Domains

Handling Controlled Terminology

Validation

- ✓ Validation tools like Pinnacle21 or OpenCDISC
- ✓ Goal is to identify and correct errors such as missing data, incorrect formats.

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

Dataset Creation

Submission Preparation

- ✓ To ensure that the mapped data adheres to regulatory standards.
- ✓ Pinnacle 21 is often used to check the accuracy

SDTM Programming

Data Collection

SDTM Mapping

Data Validation

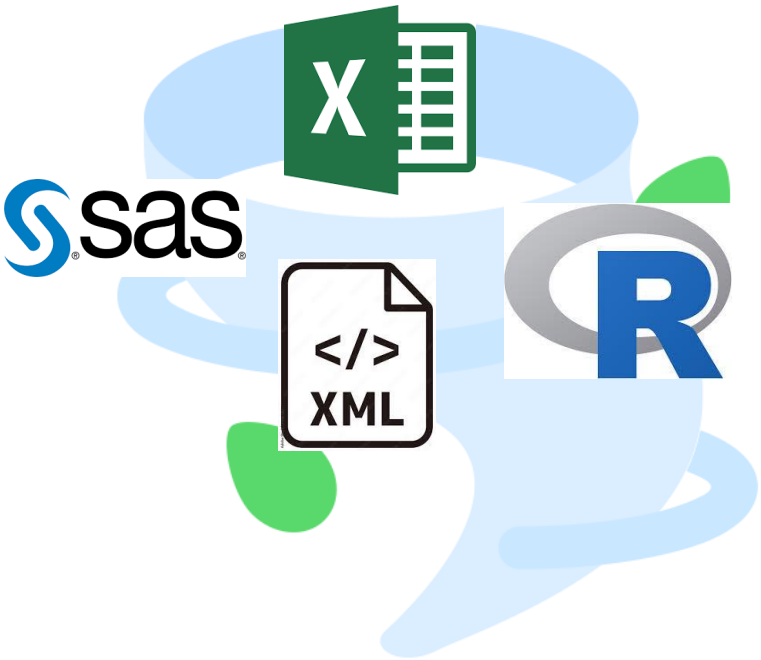
Dataset Creation

Submission Preparation

- ✓ SDTM datasets are prepared for submission to regulatory like the FDA.
- ✓ Packaging the datasets, associated documentations.
- ✓ Regulatory 제출시 XML, PDF, XPT의 확장자 파일, 가이드에서 정한 이름
- ✓ 압축은 안됨 (define.xml, XPT 파일)
- ✓ Split 폴더에 저장
- ✓ 변수, 도메인 특수문자는 ASCII

How do you convert raw data to SDTM format? (Dataset Creation)

How do you convert raw data to SDTM format?



XPT mandatory?
FDA Data Standards Catalog

FDA Data Standards Catalog v10.4

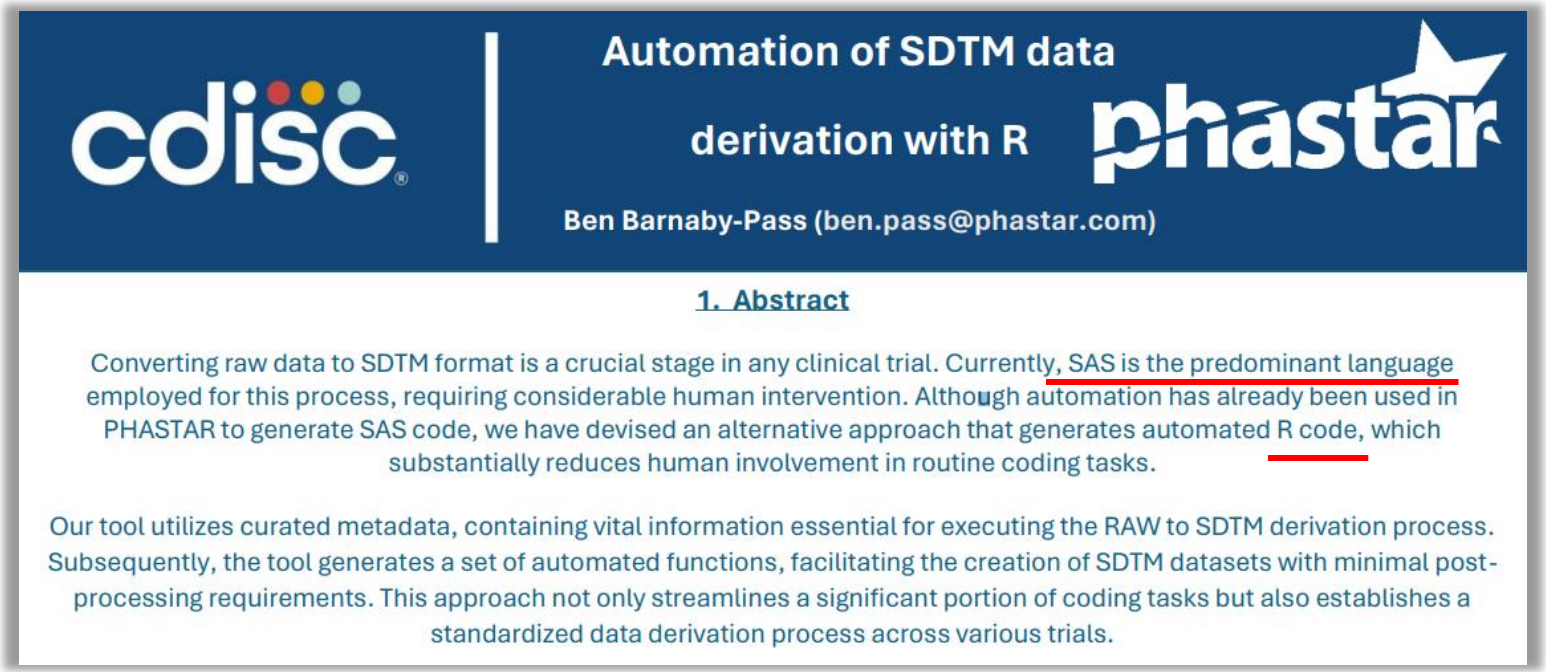
Full description of column headings in Instr.& Column Descriptions tab. Rows with data models are in bold with blue fill

Use	Standard	Exchange Format	SDO	Property
Clinical study datasets	SDTM	XPT	CDISC	SDTMIGv3.2
Clinical study datasets	SDTM	XPT	CDISC	SDTMIGv3.3
Clinical study datasets	SDTM	XPT	CDISC	SDTMIGv3.4

SAS coding

```
PROC SQL;
    CREATE TABLE EG AS
    SELECT "STUDYID" AS STUDYID
            ,"EG" AS DOMAIN
            ,T2.USUBJID AS USUBJID
            ,T1.TESTCD AS EGTESTCD
            ,T1.TEST AS EGTEST
            ,CASE WHEN T1.TESTCD="INTP" AND T1.EGRES_STD IN (2,3) THEN "ABNORMAL"
                      WHEN T1.TESTCD="INTP" AND T1.EGRES_STD=1 THEN "NORMAL" ELSE T1.ORRES END AS EGORRES
            ,CASE WHEN T1.ORRES^="" AND T1.TESTCD^="INTP" THEN "msec" ELSE "" END AS EGORRESU
            ,CALCULATED EGORRES AS EGSTRESC
            ,CASE WHEN INPUT(CALCULATED EGORRES,BEST.)^=. THEN INPUT(CALCULATED EGORRES,BEST.) END AS EGSTRESN
            ,CASE WHEN T1.EGND=0 AND T1.TESTCD^="INTP" THEN "msec" END AS EGSTRESU
            ,CASE WHEN T1.EGND=1 THEN "NOT DONE" END AS EGSTAT
            ,"12 LEAD STANDARD" AS EGMETHOD
            ,CASE WHEN T1.FOLDER2^="" AND T3.VISITNUM=2 THEN "Y"
                  WHEN T1.FOLDER2="" AND T3.VISITNUM=1 THEN "Y" END AS EGBLFL
            ,T3.VISITNUM
            ,T3.VISIT
            ,T3.VISITDY
            ,T4.EPOCH
            ,CASE WHEN T1.EGDTC_RAW^="" THEN CATX('-',PUT(T1.EGDTC_YYYY,Z4.),PUT(T1.EGDTC_MM,Z2.),PUT(T1.EGDTC_DD,Z2.)) END AS EGDTC
            ,CASE WHEN INPUT(T2.RFSTDTC,YYMMDD10.)^=. AND INPUT(CALCULATED EGDTC,YYMMDD10.)<INPUT(T2.RFSTDTC,YYMMDD10.)
                  THEN INPUT(CALCULATED EGDTC,YYMMDD10.)-INPUT(T2.RFSTDTC,YYMMDD10.)
                  WHEN INPUT(T2.RFSTDTC,YYMMDD10.) ^=. AND INPUT(CALCULATED EGDTC,YYMMDD10.)>=INPUT(T2.RFSTDTC,YYMMDD10.)
                  THEN INPUT(CALCULATED EGDTC,YYMMDD10.)-INPUT(T2.RFSTDTC,YYMMDD10.) +1 END AS EGDY
    FROM EG3 T1 LEFT JOIN SDTM.DM T2 ON T1.SUBJECT=T2.SUBJID
                        LEFT JOIN SDTM.SV T3 ON T1.SUBJECT=SUBSTR(T3.USUBJID,15,6) AND T1.VISITNUM=T3.VISITNUM
                        LEFT JOIN EPOCH_EG T4 ON T1.SUBJECT=SUBSTR(T4.USUBJID,15,6) AND T1.INSTANCENAME=T4.INSTANCENAME
;QUIT;
```

How do you convert raw data to SDTM format?



Automation of SDTM data derivation with R

Ben Barnaby-Pass (ben.pass@phastar.com)

1. Abstract

Converting raw data to SDTM format is a crucial stage in any clinical trial. Currently, SAS is the predominant language employed for this process, requiring considerable human intervention. Although automation has already been used in PHASTAR to generate SAS code, we have devised an alternative approach that generates automated R code, which substantially reduces human involvement in routine coding tasks.

Our tool utilizes curated metadata, containing vital information essential for executing the RAW to SDTM derivation process. Subsequently, the tool generates a set of automated functions, facilitating the creation of SDTM datasets with minimal post-processing requirements. This approach not only streamlines a significant portion of coding tasks but also establishes a standardized data derivation process across various trials.

Capture from CDISC

A short History of CDISC and SAS Transport Files

Background:

- When development of the SDTM and SDTMIG started, SAS was in almost universal use in the pharmaceutical industry and at FDA.
- "The SAS® Version 5 (V5) transport file format is an open standard developed by SAS to support data transfers between systems, especially those running different operating system." SAS V5, being an open standard, allowed FDA to specify is as the standard required for data submission.
- At the time, the CDISC volunteer team who developed the SDTMIG was focusing on drug submissions to FDA, so it was assumed that SDTMIG had to work with the SAS V5 transport file format. (생략).

Current Regulation:

- FDA and PMDA require SAS V5 transport file format, and other regulatory agencies accept it.
- If other recipients agreed, SDTM-based data could be exchanged in another file format. If both parties agreed, some of the restrictions rooted in the SAS V5 transport file format could be relaxed.

Changing Technology:

- Over the years, the CDISC user community has expanded, and although SAS is widely used for statistical programming in that sider community, other languages, such as R, are also used
- Non-tabular data formats, such as XML and RDF, are seeing greater use.

Text from A Short History of CDISC and SAS Transport Files | CDISC

A short History of CDISC and SAS Transport Files

Response to Changing Technology:

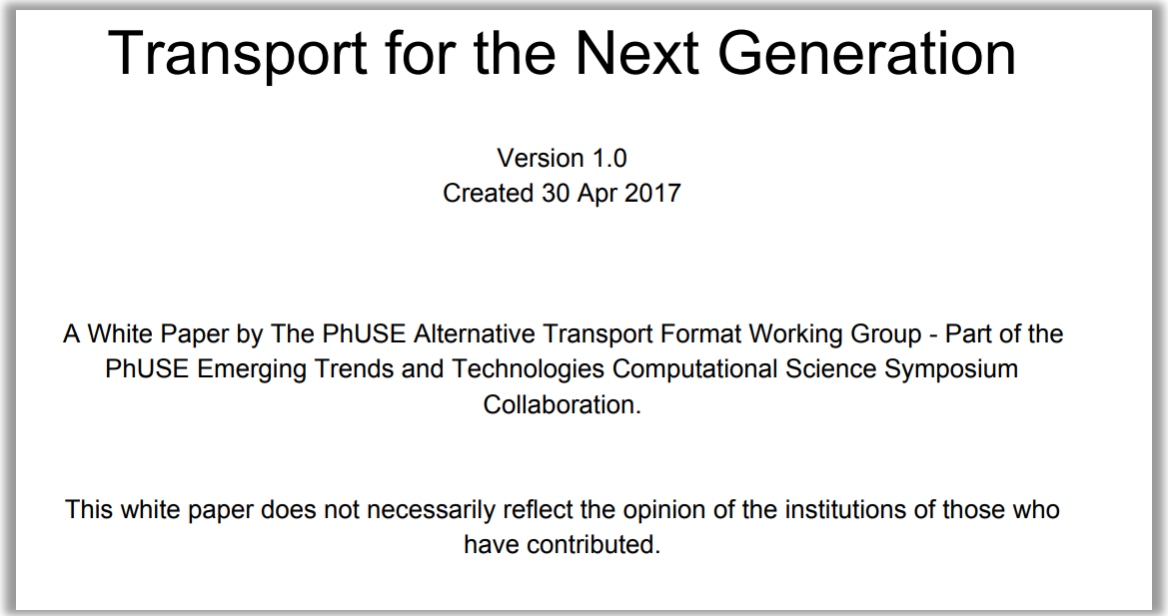
- CDISC has developed other formats, which could be used to represent SDTM-based data:
- ✓ Dataset-XML
- ✓ SDTM in RDF
- FDA has considered other formats and held a public meeting 5 Nov 2012 during which other alternatives (XML, RDF, newer SAS transport formats, etc.) were presented.
- As s result, CDISC developed Dataset-XML and SDTM in RDF and conducted some testing in collaboration with FDA.

Future Regulation:

- Regulatory agencies could require a different file format in the future
- Other parties who exchange SDTM-based data could decide to require a different exchange format. If they did, and if demand were great enough, CDISC could develop implementation advice for that different exchange format.

Text from A Short History of CDISC and SAS Transport Files | CDISC

Transport for the Next Generation



Transport for the Next Generation

Version 1.0
Created 30 Apr 2017

A White Paper by The PhUSE Alternative Transport Format Working Group - Part of the PhUSE Emerging Trends and Technologies Computational Science Symposium Collaboration.

This white paper does not necessarily reflect the opinion of the institutions of those who have contributed.

From the PHUSE Transport for the Next Generation (2017) White Paper

Transport for the Next Generation

- SAS V5 Transport format dates from 1989 and was first available as part of SAS version 5. Since that time, there have been many changes to the industry with respect to the process for submissions and the approaches to data curation and manipulation - but none to the format itself.

- SAS V5 transport format is commonly referred to as either “XPORT” (due to the LIBNAME keyword “XPORT” used during file creation) or “XPT” (due to the convention of using a file extension of “xpt”)

Pain Points of the SAS V5 transport format?

From the PHUSE Transport for the Next Generation (2017) White Paper

Dataset-JSON

CDISC and PHUSE are delighted to announce a new pilot project aimed at supporting the adoption of Dataset-JSON as an alternative transport format for regulatory submissions.

This pilot builds upon the considerable amount of work done over the years to replace XPT as the default file format for clinical and device data submissions to regulatory authorities.

…..

The pilot report will be completed in Q2 2024.

Capture from CDISC

Dataset-JSON

Dataset-JSON seeks to address the limitations of SAS V5 XPORT.

Data File Format

- Limited variable type
- Limited to US ASCII encoding
- 8-character variable names
- 40-character labels 200-character file widths

Storage

- Inefficient use of storage space
- Inability to compress datasets leads to file logistical issues
- (e.g., splitting datasets)

Content

- Lacks a robust metadata layer
- Only works for 2-dimensional data structures

Extensibility

- Not extensible

From the PHUSE Transport for the Next Generation (2017) White Paper

SAS V5 vs. V8 Format

The SAS Version 5 (V5) transport file format is an open standard developed by SAS to support data transfers between systems, especially those running different operating systems. The SAS V5 transport specification is nonproprietary.

Starting in SAS 9.3, you can use the %LOC2XPT, %XPT2LOC, and %XPTCOMMN macros to read from or write to files of V5 transport format. These macros can run equally well in SAS 8 and SAS 9. For more information, see File Transport Macros.

The V5 transport file format provides the following:

- Variable names can be up to 8 characters, and they are stored in their original case (upper or lower).
- Character variables can have lengths up to 200 bytes.
- Variable names can contain any characters other than null ('00'x).
- This transport file format allows only alphanumeric characters and underscores. Any variable name that contains characters other than alphanumeric or underscores is represented in the SAS language as an n-literal (for example, 'a b'n).

  Note: A variable name cannot be completely blank.

- Variable labels can be up to 40 characters.

From the PHUSE Transport for the Next Generation

SAS V5 vs. V8 Format

The V8 transport file format provides the following:

- Variable names can be up to 32 characters, and they are stored in their original case (upper or lower).
- Character variables can have lengths up to 32,767 bytes
- Character variable labels can have lengths up to 256 bytes.
- Variable names can contain any characters other than null ('00'x). Any variable name that contains characters other than alphanumeric or underscores is represented in the SAS language as an n-literal (for example, 'a b'n).

  Note: A variable name cannot be completely blank unless you are using the VALIDMEMNAME=EXTEND option.

- Starting in SAS 9.3, when you use the VALIDMEMNAME=EXTEND option, variable names can contain embedded blanks and these additional characters: . ~ `!@#$%^&()_+={}[];'

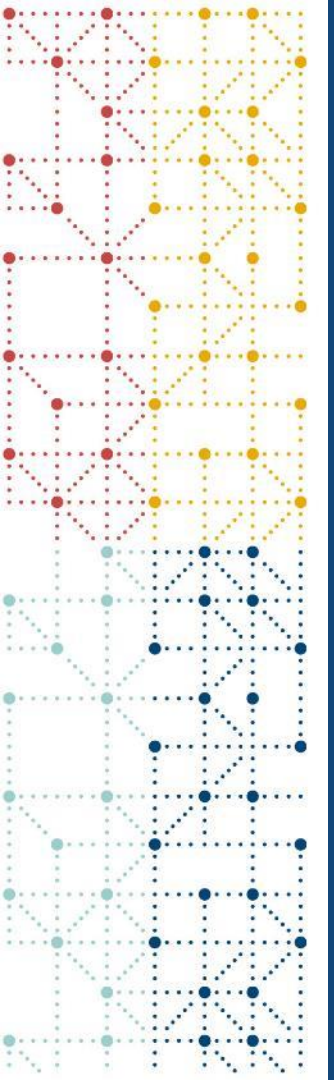
  Note: You must be running SAS 9.3 in order to use these characters in the V8 transport file format.

From the PHUSE Transport for the Next Generation

About XPT

- ✓.xpt 파일 확장자의 철자오류:
  .xpr .pt .xt .xpf .xot
- ✓.xpt 파일을 여는 데 문제:
  .xpt 파일을 열려면 해당 프로그램을 설치하기만 하면 됨.
- ✓.xpt 를 지원하는 운영체제:
  Window Windows Server 2016/2019/2022, Windows 7, Windows 8, Windows 10, Windows 11, Linux, Mac OS X, macOS, iOS, Android

From the PHUSE Transport for the Next Generation

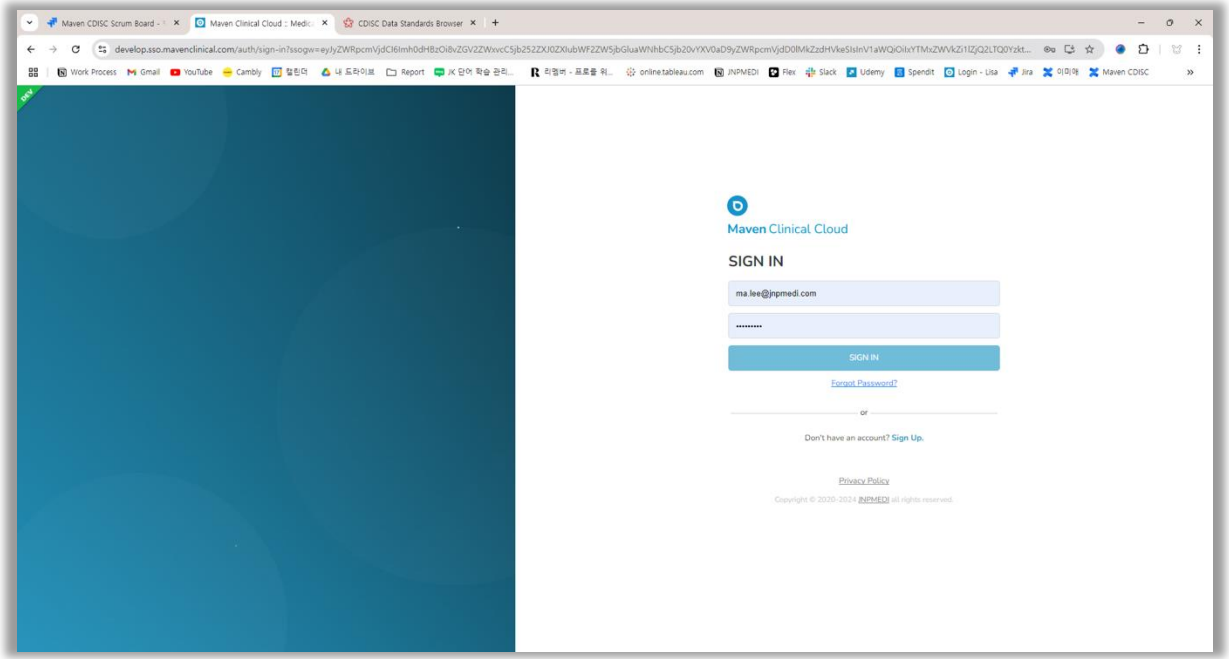


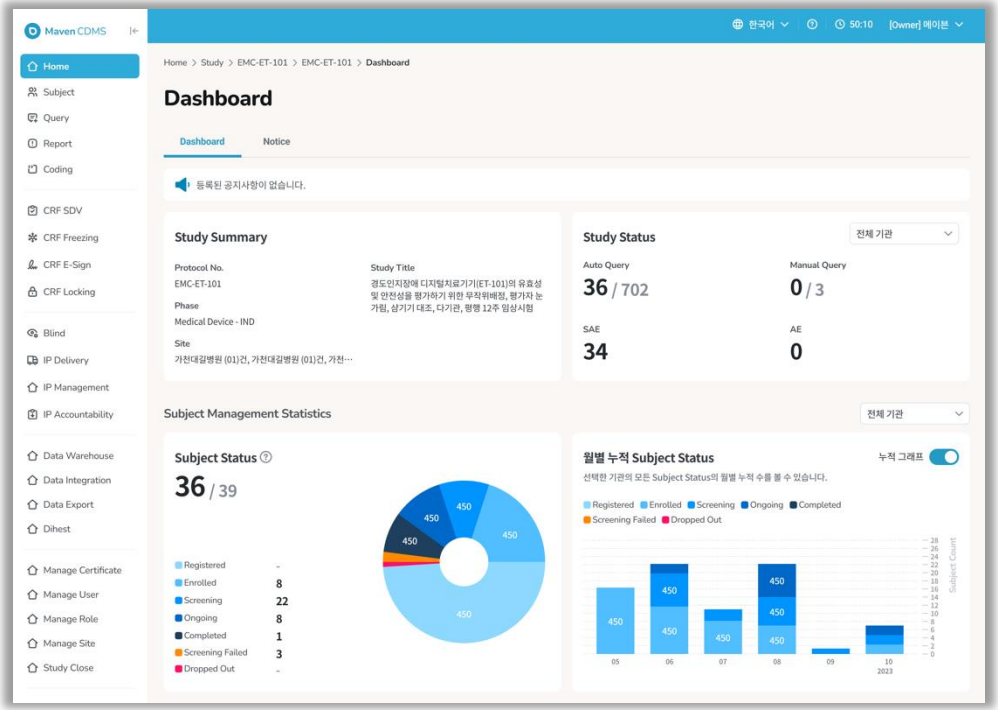
Maven Converter

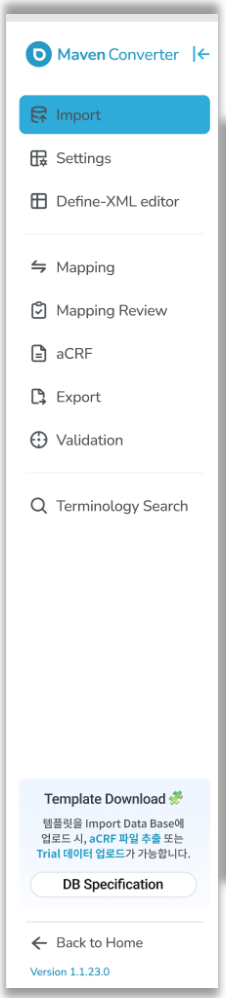
SDTM Dataset Conversion in One Click?

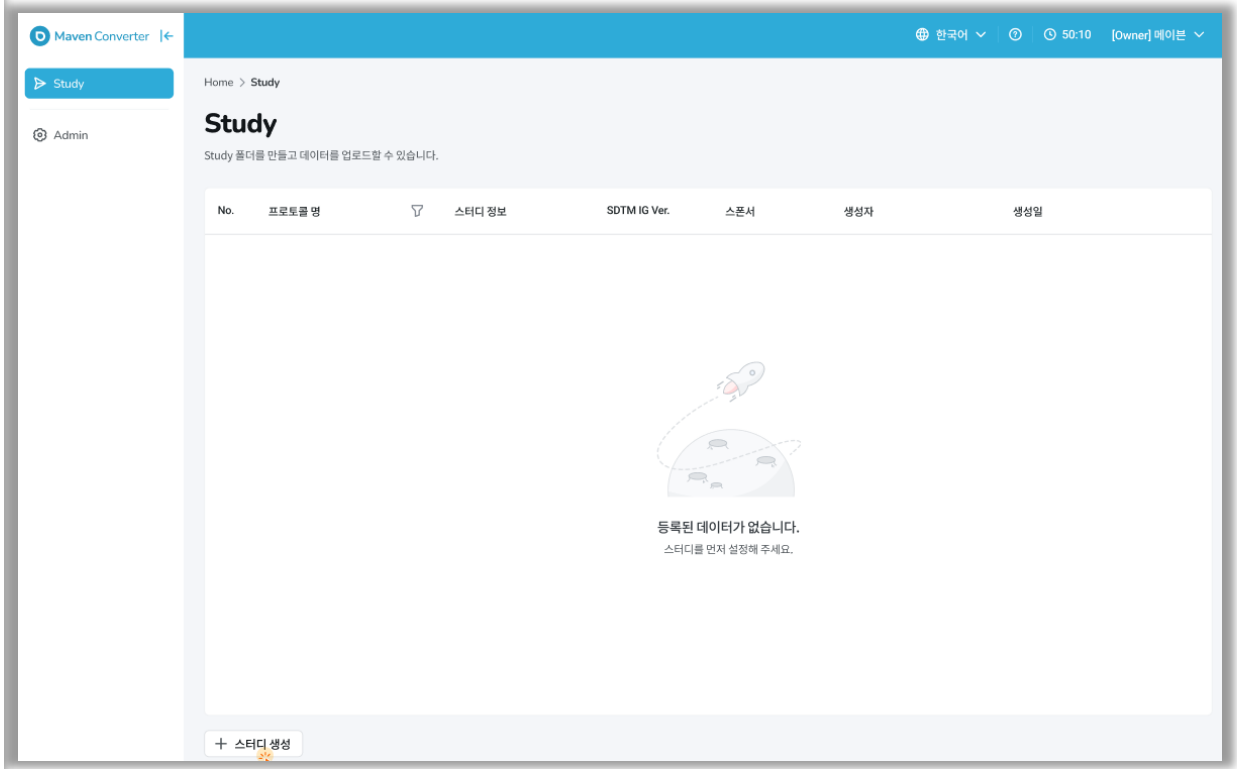
- ✓You worried that the raw datasets from your EDC need a lot of work to get them into SDTM format?
- ✓You don’t have enough time or programming skills to do it?

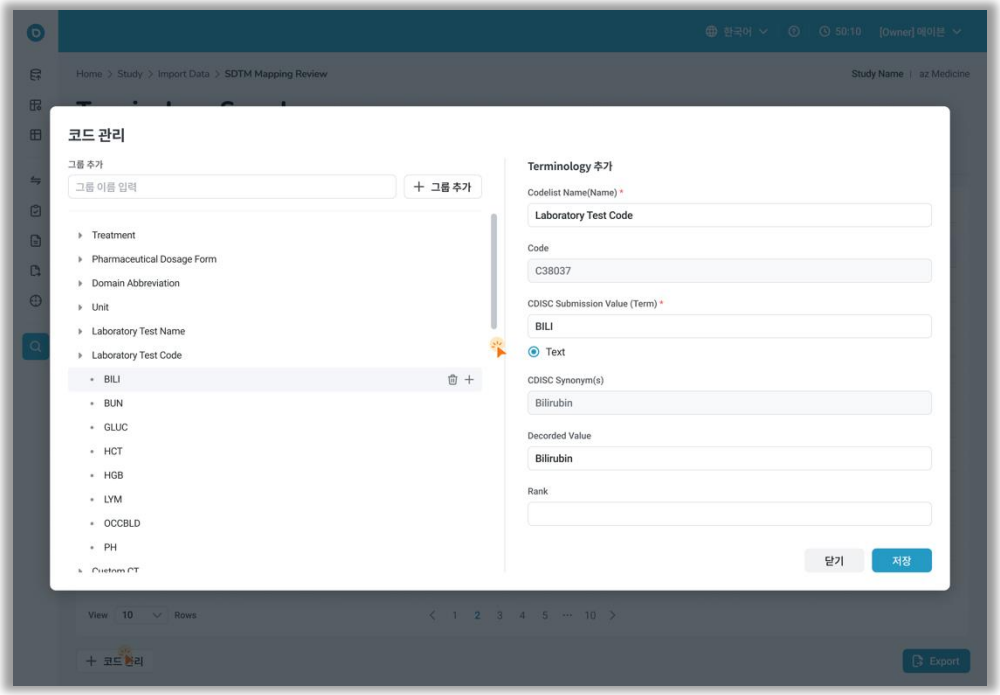
Maven Converter’s Single Sign On

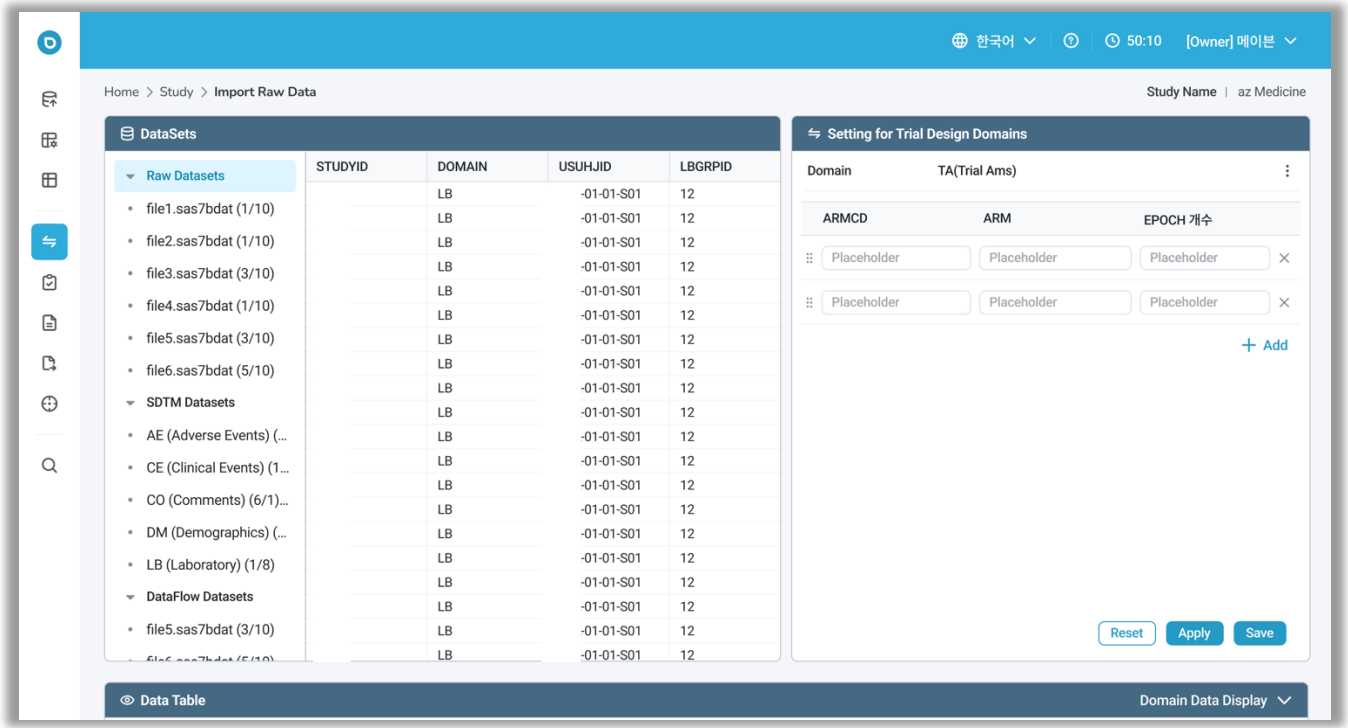


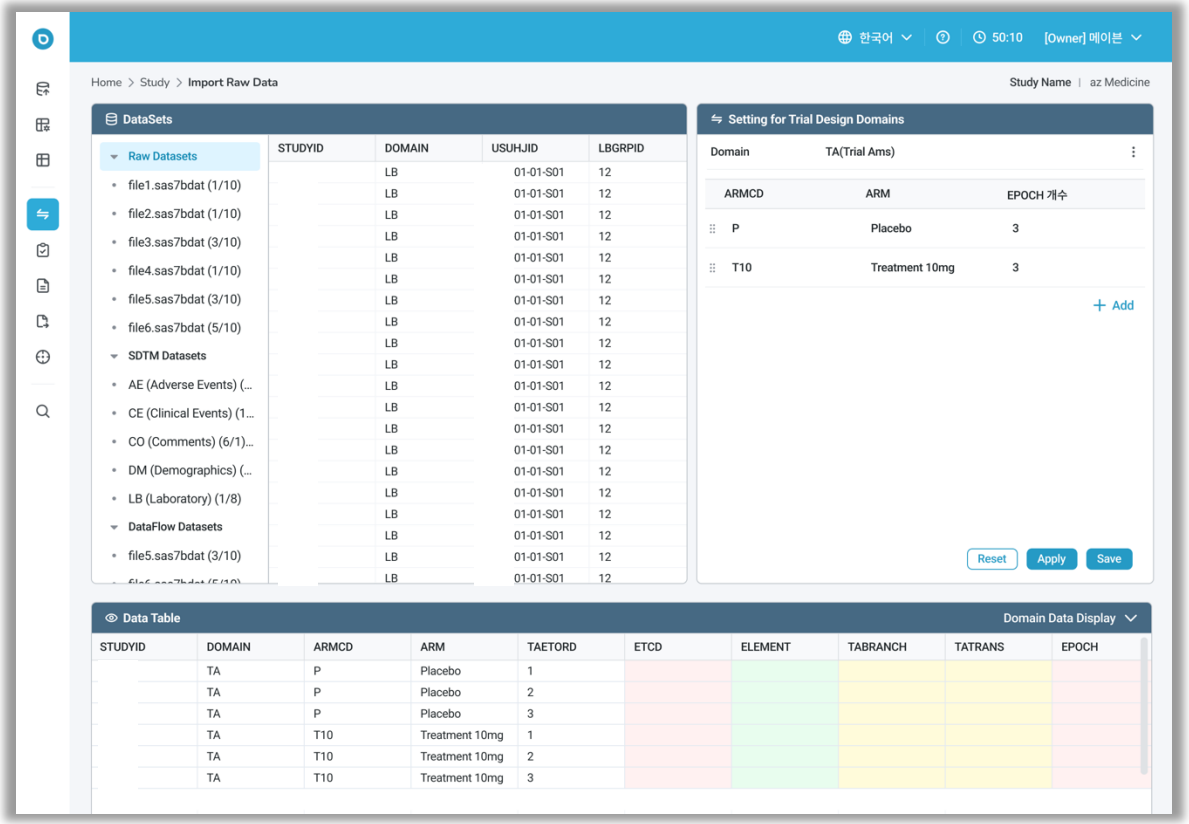

Maven CDMS
한국어
50:10
[Owner] 메이븐
Home
Subject
Query
Report
Coding
CRF SDV
CRF Freezing
CRF E-Sign
CRF Locking
Blind
IP Delivery
IP Management
IP Accountability
Data Warehouse
Data Integration
Data Export
Dihest
Manage Certificate
Manage User
Manage Role
Manage Site
Study Close
Home > Study > EMC-ET-101 > EMC-ET-101 > Dashboard
Dashboard
Dashboard
Notice
등록된 공지사항이 없습니다.
Study Summary
Protocol No.
EMC-ET-101
Phase
Medical Device - IND
Site
가천대길병원 (01)건, 가천대길병원 (01)건, 가천…
Study Title
경도인지장애 디지털치료기기(ET-101)의 유효성 및 안전성을 평가하기 위한 무작위배정, 평가자 눈가림, 삼기기 대조, 다기관, 평행 12주 임상시험
Study Status
전체 기간
Auto Query
36 / 702
Manual Query
0 / 3
SAE
34
AE
0
Subject Management Statistics
전체 기간
Subject Status
36 / 39
Registered -
Enrolled 8
Screening 22
Ongoing 8
Completed 1
Screening Failed 3
Dropped Out -
450
월별 누적 Subject Status
누적 그래프
선택한 기관의 모든 Subject Status의 월별 누적 수를 볼 수 있습니다.
Registered
Enrolled
Screening
Ongoing
Completed
Screening Failed
Dropped Out
Subject Count
05
06
07
08
09
10
2023


Maven Converter
Import
Settings
Define-XML editor
Mapping
Mapping Review
aCRF
Export
Validation
Terminology Search
Template Download
템플릿을 Import Data Base에 업로드 시, aCRF 파일 추출 또는 Trial 데이터 업로드가 가능합니다.
DB Specification
Back to Home
Version 1.1.23.0


Maven Converter
Study
Admin
한국어
50:10
[Owner] 메이븐
Home > Study
Study
Study 폴더를 만들고 데이터를 업로드할 수 있습니다.
No.
프로토콜 명
스터디 정보
SDTM IG Ver.
스폰서
생성자
생성일
등록된 데이터가 없습니다.
스터디를 먼저 설정해 주세요.
스터디 생성

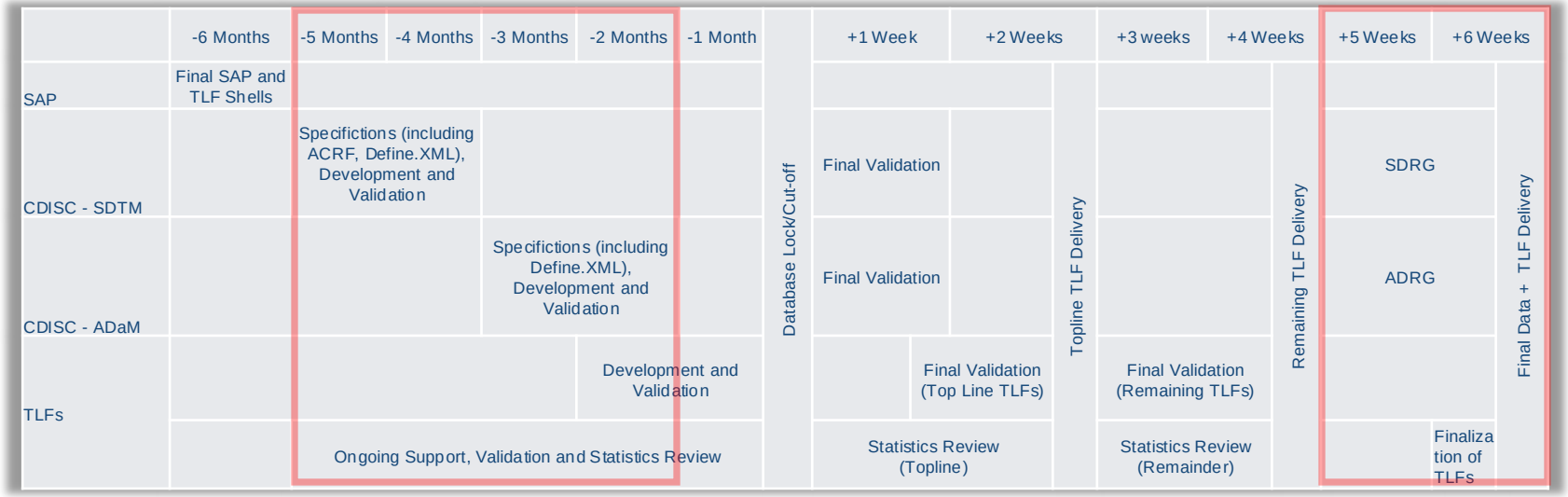

Home > Study > Import Data > SDTM Mapping Review
Study Name | az Medicine
코드 관리
그룹 추가
그룹 이름 입력
+ 그룹 추가
Treatment
Pharmaceutical Dosage Form
Domain Abbreviation
Unit
Laboratory Test Name
Laboratory Test Code
BILI
BUN
GLUC
HCT
HGB
LYM
OCCBLD
PH
Terminology 추가
Codelist Name(Name) *
Laboratory Test Code
Code
C38037
CDISC Submission Value (Term) *
BILI
Text
CDISC Synonym(s)
Bilirubin
Decorded Value
Bilirubin
Rank
닫기
저장
View 10 Rows
+ 코드 관리
Export


한국어
50:10
[Owner] 메이브
Home > Study > Import Raw Data
Study Name | az Medicine
DataSets
Raw Datasets
file1.sas7bdat (1/10)
file2.sas7bdat (1/10)
file3.sas7bdat (3/10)
file4.sas7bdat (1/10)
file5.sas7bdat (3/10)
file6.sas7bdat (5/10)
SDTM Datasets
AE (Adverse Events) (...
CE (Clinical Events) (1...
CO (Comments) (6/1)...
DM (Demographics) (...
LB (Laboratory) (1/8)
DataFlow Datasets
file5.sas7bdat (3/10)
STUDYID
DOMAIN
USUHJID
LBGRPID
LB
-01-01-S01
12
Setting for Trial Design Domains
Domain
TA(Trial Ams)
ARMCD
ARM
EPOCH 개수
Placeholder
Add
Reset
Apply
Save
Data Table
Domain Data Display

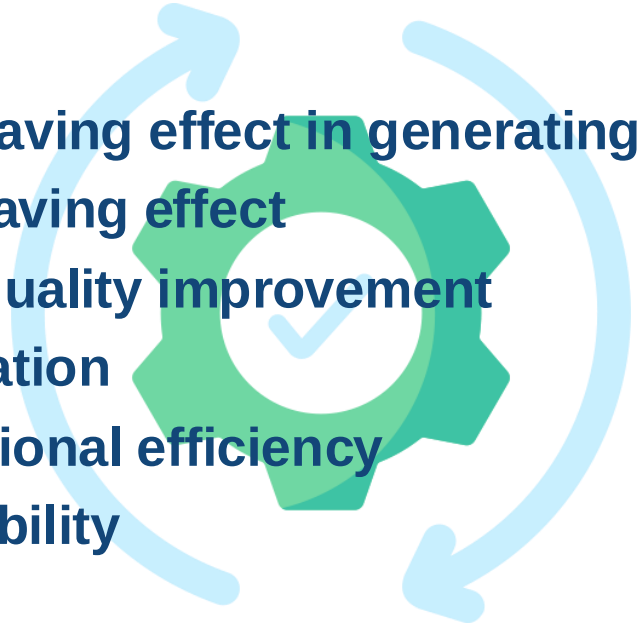


Intuitive review

Maven Converter's Expected Effect



Maven Converter's Expected Effect



- Time saving effect in generating outputs
- Cost saving effect
- Data quality improvement
- Innovation
- Operational efficiency
- Traceability

ICONS from FLATICON

Thank You!

hs.lee@jnpmedi.com