Pharma SUG 2024 - Paper SD- 318

Streamline Generation of aCRF and SDTM

Yunsheng Wang, Erik Hansen, Chao Wang, Tina Wu, ClinChoice Inc.

ABSTRACT

The manual process of annotating case report forms (CRFs) and mapping raw clinical trial data to the Study Data Tabulation Model (SDTM) standard is both resource-intensive and susceptible to human error. This traditional approach also lacks efficient traceability between CRFs, SDTM mappings files, and final SDTM datasets. This paper explores an innovative method to streamline the mapping process from Electronic Data Capture (EDC) to SDTM by utilizing the Medidata RAVE Architecture Loading Specification (ALS) file. The ALS can be customized to establish links between EDC specifications, CRF, SDTM mapping specifications, and SDTM datasets, creating a seamless end-to-end data flow. By leveraging the customized ALS file, this innovative approach enhances traceability, accuracy, and overall efficiency compared to traditional manual methods. The demonstration of the automated mapping process covers the entire spectrum from EDC data collection to the generation of SDTM define XML. Serving as a potential model for optimizing EDC to SDTM mapping workflows, this structured and automated approach reduces the dependence on manual work, therefore improving data quality and expediting drug development timelines.

INTRODUCTION

Traditionally, the standard approach to SDTM mapping begins with annotating CRFs based on SDTM Implementation Guide (IG), followed by creating the SDTM mapping specification to clarify the conversion definition between collected data and SDTM variables. Once reviewed, development and validation programmers follow the SDTM mapping specification to program and generate final SDTM datasets. Upon completion of these steps, the define package can be developed. However, this method requires multiple steps and manual adjustments, particularly when there are updates in the CRF. Manual adjustments involve modifying aCRF, SDTM mapping files, and SDTM programs, making this conventional method inefficient. Considering the clinical trial may span an extended period ranging from 5 to 10 years, during which the protocol and case report forms may undergo updates ranging from 5 to 15 times, it becomes particularly challenging. With each update, the traditional approach struggles to effectively capture and maintain traceability, as well as ensure accuracy across all elements.

On the contrary, the automated ALS to SDTM process depicted in **Figure 1**, significantly reduces the need for manual intervention while also maintaining traceability and accuracy across different updates. With the development of our company's automation tools, and our dynamically updated Knowledge Banks, Clinchoice has significantly streamlined many manual tasks. The ALS file encompasses various components such as forms, fields, data dictionaries, control types, and more, all of which are essential parameters for mapping our final SDTMs. Utilizing this comprehensive document enables us to streamline the generation of annotated CRFs, SDTM mapping specifiations, SDTMs, and Define XML.

The ALS and Codelist Knowledge Banks have been meticulously crafted from thoroughly reviewed related studies that have successfully passed all P21 checks. This guarantees that the new study maintains consistency with previously submitted studies. Additionally, the Knowledge Bank will continuously update based on the latest reviewed and submitted studies. This ensures the annotation and codelist formats extracted from the Knowledge Bank remain accurate. Moreover, the Knowledge Bank will continue to expand as additional studies are included. This expansion will enhance the power of the automation process, diminish the need for manual intervention, and ultimately enhance the efficiency of integrating new studies.

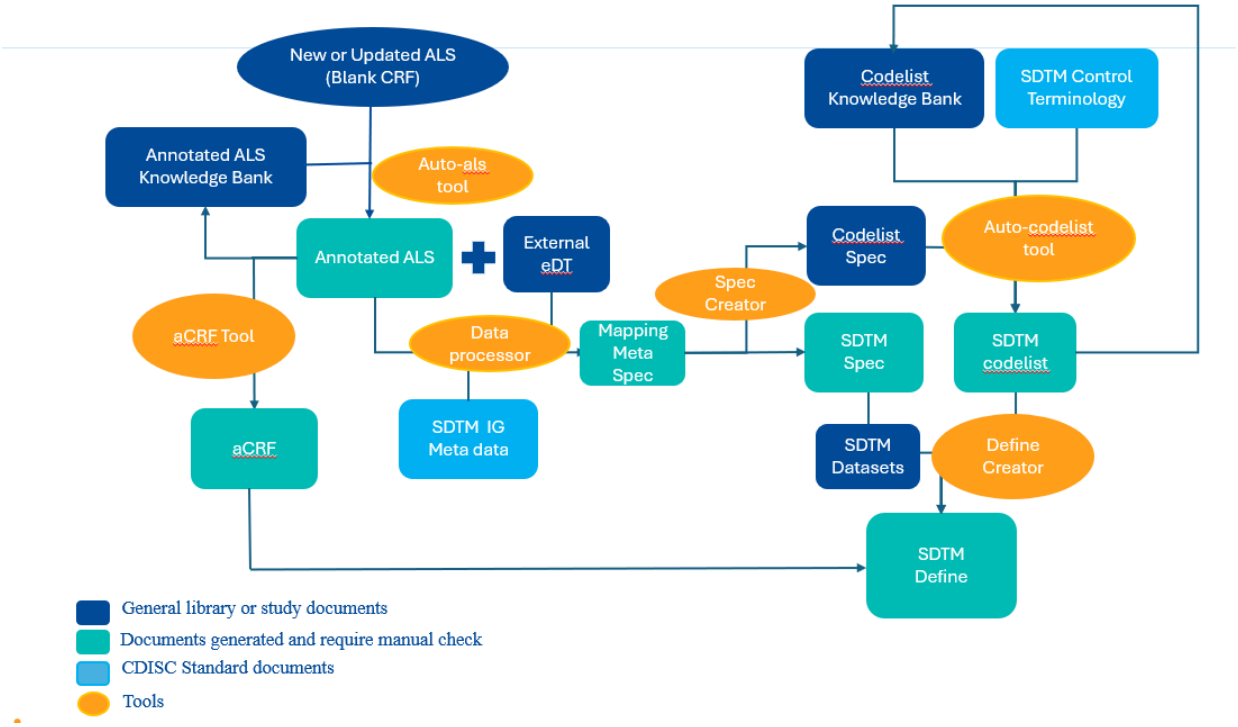


Figure 1. Process of generation of aCRF and SDTM packages

AUTO GENERATING ANNOTATED ALS AND ANNOTATED CRF

Utilizing the annotated ALS Knowledge Bank, when commencing a study, we employ our auto-annotated ALS tool to search the Knowledge Bank for similar CRF forms. The Auto-Annotated ALS Program focuses on finding 'acrf_text' values for both the Forms and Fields from ALS Meta Specifications. These 'acrf_text' values are crucial to creating an accurate Annotated Case Report Form Document, which is an invaluable reference to our developers throughout the SDTM process.

The annotated ALS automation is designed to reduce the manual efforts required by developers in mapping ALS Meta Specifications. Thus, a thorough and systematic approach is adopted to ensure the precision and reliability of the tool's outputs. In the new study's ALS, the FormOID, FieldOID, and pretext are utilized to locate a match for FieldOID, FormOID, and pretext in the ALS Knowledge Bank. Subsequently, the corresponding 'acrf_text' is assigned to the new study's ALS (see **Display 1**).

acrf_text	OID
PR;CO	MRI_CT1
BE	BIOP_ONSTUDY
BE;MI	BIOP_HX
BE;MI	BIOP_SCR
EC;EX;FI;CO	EX_LOG
FE;GF	LFSTYL
PR;ZI	FS
VS;CO	VS
EG	EG
PR;LB;CO	LB_URINE
	LB_URIN2
PR	LB_NSC
PR	LB_BLOOD
PR	LB_BLOOD_UNS
LB;PR;CO	LB_LOCAL
NOT SUBMITTED	RESCR
PR;PC;CO	PK_TRO
PR;CO	BMD
LB;PR;RP;CO	LB_REPSYS
PR;CO	LB_GENSAM
PR;CO	LB_BIOM
Not SUBMITTED	BLT
FI	INTRUP
DS;DM	EOT
DS;DM	EOT_EXT
DS;DM	EOS

Forms | Fields

acrf_text	FormOID	FieldOID
NOT SUBMITTED	SUBJC	CCG
SUPPDM.QVAL where QNAM = RACEOTH	SUBJC	RTSMML
NOT SUBMITTED	SUBJC	SCRYN
DM.COUNTRY	SUBJC	COUNTRY
NOT SUBMITTED	SUBJC	SUBCRYN
DM.BRTHDTC	SUBJC	PBRTHYY
DM.RACE, when multiple values are select	SUBJC	SCRDAT
	SUBJC	SUBPRYN
	SUBJC	SUBPRID
DM.RACE, when multiple values are select	SUBJC	LABEL1
DM.STUDYID	SUBJC	STUDYID
DM.SITEID	SUBJC	SITEID
NOT SUBMITTED	SUBJC	SUBNUM
DM.SUBJID, DM.USUBJID	SUBJC	USUBJID
NOT SUBMITTED	SUBJC	SUBSTAT
NOT SUBMITTED	SUBJC	SUBDATE
NOT SUBMITTED	SUBJC	SUBDATRD
VE.VEOCCUR	SV	VISYN
VE.VETERM	SV	VISTYP
VE.VEDTC;SV.SVSTDTC, SV.SVENDTC	SV	VISDAT
VE.VEREASOC	SV	VSREASND
SUPPVE.QVAL where QNAM = VERESOTH	SV	VSRESOTH
SUPPSV.QVAL where QNAM = NWSITEYN	SV	NWSITEYN
SUPPSV.QVAL where QNAM = NWSITEID	SV	NWSITEID
DS.DSTERM	ICF	FICFYN
DS.DSSTDTC	ICF	FICDAT

Forms | Fields

Display 1. Auto generated annotated ALS Form and Field sheet with acr_text

The automatically annotated ALS ‘acrf_text’ will undergo manual review and updates by the programming team. This process ensures alignment with the protocol, CRF, and the latest SDTM IG, guaranteeing that the final Annotated ALS complies with the study and the most current CDISC standards. The Annotated ALS includes the raw dataset name, raw variable name, raw data dictionary codelist, and their corresponding mapped SDTM dataset name and variable name. This file is instrumental in creating a study-customized codelist format and mapping specifications for programming and delivery package preparation purposes.

With the reviewed Annotated ALS, the aCRF tool creates the aCRF (**Display 2.**).

DM=Demographics

Version 1.14 PPC14: UNIQUE
Project Name: [redacted]
Form: Demographics
Generated On: 27 Jul 2023 16:13:14

Year of Birth DM.BRTHDTC
Collected Age DM.AGE
Collected Age unit DM.AGEU Years
Age is based on the first informed consent date
Age in Years (Derived) NOT SUBMITTED
Sex at Birth DM.SEX Female ☐ Male ☐
Ethnicity DM.ETHNIC Hispanic or Latino ☐ Not Hispanic or Latino ☐ Not Reported ☐

Race (Check all that apply)
American Indian or Alaska Native SUPPDM.QVAL where QNAM = RACE1
Asian SUPPDM.QVAL where QNAM = RACE2
Black or African American SUPPDM.QVAL where QNAM = RACE3
Native Hawaiian or Other Pacific Islander SUPPDM.QVAL where QNAM = RACE4
White SUPPDM.QVAL where QNAM = RACE5
Not Reported DM.RACE, when multiple values are selected then RACE=MULTIPLE
Unknown

Display 2. Automatically generated aCRF

MAPPING SPECIFICATION

Following the creation of the Annotated ALS and aCRF, it is important to format the external data transfer information in a similar format to that of the Annotated ALS. This step is crucial to capture both the Annotated ALS and the external data information in the mapping specifications. The Data Processor Program generates the necessary dictionaries and structures the eDT data to align with the reviewed Annotated ALS from the previous step efficiently. Subsequently, these components are processed together to create the SDTM and Codelist Mapping Meta Specification. The mapping specifications are generated based on the versions of ALS (CRF) and data transfer. It labels the mapping specifications with the most recent date, referred to as the data transfer date, ensures all updates between different iterations and runs are captured and documented. This feature enhances traceability and accuracy throughout the duration of the long-term clinical trial.

MAPPING META SPECIFICATION

The generated Mapping Meta Specification, illustrated in **Display 2**, includes Formoid (collected and external raw dataset names), Fieldoid (collected and external raw variable names), aCRF_text (collected raw data mapping annotation based on SDTM IG), domain (standard mapping SDTM domain names), variable (standard mapping SDTM variable names), dsource (source from external or CRF collected), and cd_: (the assigned codelist for the mapping variables).

The domain, variable, and cd_: values are parsed from the aCRF_text using the Annotated ALS file. For external data, domain and variable assignments are derived from the external mapping specification. Any necessary updates to the domain, variable, and cd_: can be made in this file, and the saved file can be executed to ensure that these updates are reflected in future runs and revisions.

	A	B	C	D	E	F	G	H	I	J	K
1	aCRF_text	fieldoid	formoid	domain	variable	cd_1	cd_2	cd_3	cd_4	cd_5	dsource
52		EGBLFL	CECG	EG	EGBLFL						External
53		EGCAT	CECG	EG	EGCAT						External
54		EGDTC	CECG	EG	EGDTC						External
55		EGEVAL	CECG	EG	EGEVAL						External
56		EGLEAD	CECG	EG	EGLEAD						External
57		EGMETHOD	CECG	EG	EGMETHOD						External
58		EGNAM	CECG	EG	EGNAM						External
59		EGORRES	CECG	EG	EGORRES						External
60		EGORRESU	CECG	EG	EGORRESU						External
61		EGPOS	CECG	EG	EGPOS						External
62		EGREFID	CECG	EG	EGREFID						External
63		EGSEQ	CECG	EG	EGSEQ						External
64		EGSTRESC	CECG	EG	EGSTRESC						External
65		EGSTRESN	CECG	EG	EGSTRESN						External
66		EGSTRESU	CECG	EG	EGSTRESU						External
67		EGTEST	CECG	EG	EGTEST						External
68		EGTESTCD	CECG	EG	EGTESTCD						External
69		USUBJID	CECG	EG	USUBJID						External
70		VISIT	CECG	EG	VISIT						External
71		VISITNUM	CECG	EG	VISITNUM						External
107	SUPPEG.QVAL where QNAM = EGEPRELI	EGCVDYN	EG	SUPPEG	EGEPRELI						CRF
108	EG.EGDTC	EGDAT	EG	EG	EGDTC						CRF
109	EG.EGSTAT	EGOCCUR	EG	EG	EGSTAT						CRF
110	EG.EGORRES where EG.EGTESTCD=INTP	EGOVRES	EG	EG|EG	EGORRES|EGTESTCD		INTP				CRF
111	SUPPEG.QVAL where QNAM = EGPDC	EGPDC	EG	SUPPEG	EGPDC						CRF
112	EG.EGREASND	EGREASOC	EG	EG	EGREASND						CRF

Display 2. ALS to SDTM mapping specification

If there is an update to the Annotated ALS, or a new data transfer, then re-run the Data Processor and a new Mapping Meta Specification will be generated which will capture the previous updates and incorporate them into current updates. Additionally, it will flag (Flag='Y') and highlight (peach puff) the differences between the current and previous data transfers, as demonstrated in **Display 3**. Now,it is clear to reviewers that a new test result has been added to MI. Therefore, when reviewing SDTM datasets, particular attention should be given to this MI domain.

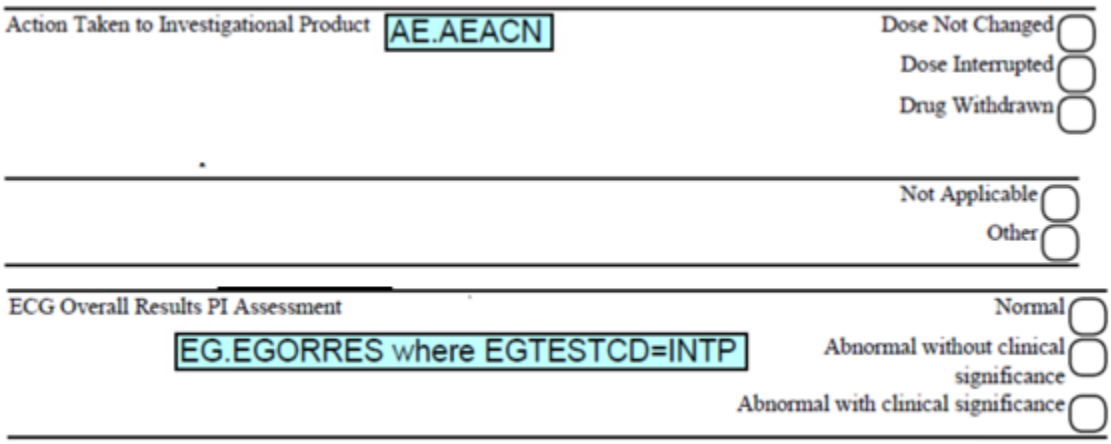
1	aCRF_text	fieldoid	formoid	domain	variable	cd_1	cd_2	cd_3	cd_4	cd_5	dsource	flag
11	AE.AESDISAB	AESDISAB	AE	AE	AESDISAB						CRF	
12	AE.AESDTH	AESDTH	AE	AE	AESDTH						CRF	
13	AE.AESER	AESER	AE	AE	AESER						CRF	
14	AE.AESHOSP	AESHOSP	AE	AE	AESHOSP						CRF	
15	AE.AESLIFE	AESLIFE	AE	AE	AESLIFE						CRF	
16	AE.AESMIE	AESMIE	AE	AE	AESMIE						CRF	
17	AE.AESTDTC	AESTDAT	AE	AE	AESTDTC						CRF	
18	AE.AESTDTC	AESTTIM	AE	AE	AESTDTC						CRF	
19	AE.AETERM	AETERM	AE	AE	AETERM						CRF	
20	AE.AETOXGR	AETOXGR	AE	AE	AETOXGR						CRF	
21	SUPPAE.QVAL where QN	DTHDAT	AE	SUPPAE|DM	AEDTHDTC|DTHDTC						CRF	
22	HO.HOENDTC	HOENDAT	AE	HO	HOENDTC						CRF	
23	HO.HOENDTC	HOENTIM	AE	HO	HOENDTC						CRF	
24	HO.HOSTDTC	HOSTDAT	AE	HO	HOSTDTC						CRF	
25	HO.HOSTDTC	HOSTTIM	AE	HO	HOSTDTC						CRF	
26	MI.MITESTCD=FIBROSIS and	BIOFIBH	BIOP_HX	MI|MI|MI|MI	MIORRES|MITESTCD|MISTDTL|MICA		FIBROSIS	NASH CRN	CENTRAL READER		CRF	Y
27	MI.MITESTCD=NAS and MI.MITSTDTL=BALLOO	BIOHXBDS	BIOP_HX	MI|MI|MI	MIORRES|MITESTCD|MITSTDTL		NAS	BALLOONING			CRF	Y
28	MI.MIDTC; BE.BESTDTC	BIOHXDAT	BIOP_HX	MI|BE	MIDTC|BESTDTC						CRF	
29	MI.MITESTCD=FIBROSIS and	BIOHXFS	BIOP_HX	MI|MI|MI	MIORRES|MITESTCD|MISTDTL		FIBROSIS	NASH CRN			CRF	Y
30	MI.MITESTCD=FIBROSIS and MI.	BIOHXFSO	BIOP_HX	MI|MI|MI	MIORRES|MITESTCD|MISTDTL		FIBROSIS	NASH CRN			CRF	Y
31	MI.MITSTDTL	BIOHXFSS	BIOP_HX	MI	MITSTDTL						CRF	
32	MI.MITESTCD=NAS and MI.MITSTDTL=LOBULA	BIOHXLIS	BIOP_HX	MI|MI|MI	MIORRES|MITESTCD|MITSTDTL		NAS	LOBULAR INFLAMMATION			CRF	Y

Display 3. ALS to SDTM mapping Meta Specification with differences highlighted and flagged between various runs.

MAPPING META CODELIST

The creation of the Codelist Meta Specification is contingent upon the ALS control type, dictionary, and 'acrf_text' column. If the data dictionary name provided in the ALS file indicates that the collected CRF has a dropdown list, or if the annotation has assigned values (EGTESTCD=INTP), as depicted in **Display 4**, those variable values should adhere to SDTM Control Terminology (CT) per FDA define submission package requirements (Evgeny Starostin, 2019). It also incorporates raw frequency values of external data if the mapped SDTM variable necessitates a format as per the SDTM IG.

Utilizing the Annotated ALS, the domain and variable names can be extracted from the 'acrf_text' column. Subsequently, these domain and variable names can be merged with SDTM IG metadata to obtain standard CT names. As illustrated in **Display 5**, the produced Codelist Meta Specification includes the raw codelist values, final SDTM dataset name and variable name. Additionally, it features CTNAME, indicating which SDTM variables require proper formatting during programming and must be accurately displayed in the define xml. It also captures, highlights, and flags any updates and discrepancies between different runs. The reviewed Codelist Meta Specification can then be used to auto generate study customized format and Codelist.



Display 4. aCRF with dropdown list

1	DATASE	NAME	DATADICTIONARYNAME	CTNAME	RAW	edtmappe	dsource	flag
2	AE	AEACN	AE Action Taken	ACN	Dose Interrupted		CRF	
3	AE	AEACN	AE Action Taken	ACN	Dose Not Changed		CRF	
4	AE	AEACN	AE Action Taken	ACN	Drug Withdrawn		CRF	
5	AE	AEACN	AE Action Taken	ACN	Not Applicable		CRF	
6	AE	AEACN	AE Action Taken	ACN	Other		CRF	
55	CM	CMDOSU	CM Dose Units	UNIT	U = Unit		CRF	Y
56	CM	CMDOSU	CM Dose Units	UNIT	cap = Capsule		CRF	
57	CM	CMDOSU	CM Dose Units	UNIT	g = Gram		CRF	
58	CM	CMDOSU	CM Dose Units	UNIT	gtt = Drop		CRF	
59	CM	CMDOSU	CM Dose Units	UNIT	mL = Milliliter		CRF	
60	CM	CMDOSU	CM Dose Units	UNIT	mcg = Microgram		CRF	
61	CM	CMDOSU	CM Dose Units	UNIT	mg = Milligram		CRF	
62	CM	CMDOSU	CM Dose Units	UNIT	tab = Tablet		CRF	
63	CM	CMDOSU	CM Dose Units	UNIT	tsp = Teaspoon		CRF	
182	EG	EGMETHOD		EGMETHOD	12 LEAD STANDARD		External	
184	EG	EGORRES	Clinical Significance	EGORRES	Abnormal with clinical significance		CRF	
185	EG	EGORRES	Clinical Significance	EGORRES	Abnormal without clinical significance		CRF	
186	EG	EGORRES	Clinical Significance	EGORRES	Normal		CRF	
212	EG	EGTESTCD		EGTESTCD	AVCOND	Atrioventricula	External	
213	EG	EGTESTCD		EGTESTCD	AXISVOLT	Axis and Volta	External	
214	EG	EGTESTCD		EGTESTCD	CHYPTENL	Hypertrophy or	External	
215	EG	EGTESTCD		EGTESTCD	EGHRMN	ECG Mean H	External	
216	EG	EGTESTCD		EGTESTCD	INTP		CRF	

Display 5. Codelist meta specification

SDTM CODELIST AUTO MAPPING:

Central to this model is the accurate mapping of raw data elements to standardized CDISC submission values, a process traditionally prone to manual effort and potential errors.

METHODOLOGY:

The Codelist Automapping program utilizes the Codelist Meta Specifications (reference above for more information), the SDTM Controlled Terminology, and a ClinChoice SDTM Codelist Knowledge Bank to map raw data elements to standardized CDISC submission values. The program operates in a structured manner, leveraging the following steps:

1. **Initial Check:**
   The program scans the Codelist Meta Specification sheet to determine Controlled Terminology Name (CTNAME) associated with each raw value.
2. **Direct Mapping:**
   If the CTNAME exists in the SDTM Controlled Terminology (CT), the program checks for a direct mapping to a CDISC submission value. If found, the mapping is stored in the output codelist map.
3. **Historical Mapping Retrieval:**
   In instances where direct mapping is unavailable, the program references the SDTM Codelist Knowledge Bank (reference section below for more information) to retrieve historical mappings from prior studies. If a match is found, it is included in the output Codelist map.
4. **Intelligent Guesswork**:
   When neither direct mapping nor historical mappings are available, the program employs the Python Library known as Fuzzy Wuzzy to identify the top three best matches from the SDTM CT Excel, based on similarity scores.
   a. The Python Fuzzy Wuzzy Package employs the Levenshtein Distance algorithm to determine a similarity score between two strings.
   b. The SDTM Codelist Automapping Program derives a similarity score between the Raw Value found in the Codelist Meta Specifications and the following components:

i. CDISC Submission Value

ii. CDISC Synonyms

iii. CDISC Definition

iv. NCI Preferred Term

c. Using a sum of all these scores with preferential scaling given to the CDISC Submission Value, the Auto Mapping Program can determine a row similarity score. The top three best CDISC Submission Value matches from the SDTM Controlled Terminology are then output to the Codelist Automap Excel, providing multiple available options during manual review.

THE CODELIST AUTO MAP OUTPUT:

The Codelist Auto Mapping program generates an output consisting of two sheets, each serving distinct mapping purposes. This section provides an overview of the structure and functionality of these output sheets.

1. **Test Name and Test Code Mapping Sheet:**
   The first sheet of the output is dedicated to mapping raw values corresponding to Test Names or Test Codes. In the SDTM framework, each Test Name has a corresponding Test Code, and vice versa. Therefore, this sheet facilitates mapping in a 1:2 manner, wherein a single raw value input yields two corresponding CDISC Submission Values: one for the Test Name and another for the Test Code. Reference **Display 6** for a visual of the 1:2 mapping sheet.

F	G	H	I	J	K	L	M
		Direct Map or First Best Guess		Second Best Guess		Third Best Guess	
RAW	EDTMAPPED	TEST 1	TESTCD 1	TEST 2	TESTCD 2	TEST 3	TESTCD 3
ALT		Alanine Aminotransferase	ALT				
APTT		Activated Partial Thromboplastin Time	APTT				
AST		Aspartate Aminotransferase	AST				
Adiponectin		Adiponectin	ADPNCTN				
Albumin		Albumin	ALB				
Alkaline phosphatase		Alkaline Phosphatase	ALP				
Anti-HBc (Hepatitis B)		Hepatitis B Virus Core Antigen	HBCAG	Hepatitis B Virus e Antigen	HBEAG	Hepatitis C Virus Core Antigen	HCCAG
Anti-HBs (Hepatitis B)		Hepatitis B Virus e Antigen	HBEAG	Hepatitis B Virus Core Antigen	HBCAG	Hepatitis C Virus Core Antigen	HCCAG
Anti-HCV (Hepatitis C)		Hepatitis C Virus Core Antigen	HCCAG	Hepatitis B Virus Core Antigen	HBCAG	Hepatitis C Virus Antigen	HCAG
Basophils		Basophils	BASO				
Basophils (abs.)		Basophils	BASO				
Bicarbonate		Bicarbonate	BICARB				
Bilirubin, Direct		Direct Bilirubin	BILDIR				
Bilirubin, total		Bilirubin	BILI	Bilirubin Crystals	CYBILI	Direct Bilirubin/Bilirubin	BILDIRBI
Blood urea nitrogen		Urea Nitrogen	UREAN				
C-peptide		C-peptide	CPEPTIDE				

Display 6: Test/Test Code Codelist Automap Sheet

2. **Non-Test/Non-Test Code Mapping Sheet:**
   The second sheet of the output caters to mapping non-Test and non-Test Code values. Unlike the first sheet, mapping in this context follows a 1:1 relationship, where a single raw value input corresponds directly to a CDISC Submission Value. Reference **Display 7** for a visual of the 1:1 mapping sheet.

		Direct Map or Best Guess
CTNAME	RAW	MAPPED
UNIT	gtt = Drop	gtt
UNIT	mL = Milliliter	mL
UNIT	mcg = Microgram	ug
UNIT	mg = Milligram	mg
UNIT	tab = Tablet	TABLET
UNIT	tsp = Teaspoon	tsp
UNIT	beats/min	beats/min
UNIT	msec	ms
UNIT	%	%
UNIT	/uL	10^6/L
UNIT	10^3/uL	10^9/L
UNIT	10^9/L	10^9/L
UNIT	IU/L	IU/L
UNIT	U/L	U/L
UNIT	g/L	g/L
UNIT	g/dL	g/L
UNIT	mEq/L	mEq/L
UNIT	mL/min	mL/min
UNIT	mg/100mL	dpm/100 mg
UNIT	mg/L	mg/L
UNIT	mg/dL	mmol/L
UNIT	mg/mL	g/L

Display 7: Non-Test Mapping Sheet

THE SDTM CODELIST KNOWLEDGE BANK AND ITS UPDATE PROCESS:

Following the execution of the Codelist Automapping program, a systematic procedure is employed to update the SDTM Codelist Knowledge Bank, ensuring its alignment with the most current mappings. This process entails the following steps:

1. **Identification of Derived Mappings:**
   Upon completion of the Codelist Auto Mapping process, rows within the output Auto Mapped Codelist that obtained mappings from prior mappings stored in the SDTM Codelist Knowledge Bank or through intelligent guesswork are highlighted in yellow. This visual cue prompts programmers to prioritize the review of these specific rows.
2. **Review and Validation:**
   Programmers meticulously review the highlighted rows to verify and validate the derived mappings, ensuring their accuracy and compliance with CDISC standards. Any discrepancies or inaccuracies are addressed and rectified during this review phase.
3. **Execution of Update SDTM Codelist Knowledge Bank Program:**
   Subsequently, the Update SDTM Codelist Knowledge Bank program is deployed to facilitate the systematic update of the SDTM Codelist Knowledge Bank. This Python-based program operates as follows:
   a. *Scanning and Comparison:*
      The program scans through the current Knowledge Bank, identifying existing mappings between raw values and CDISC Submission Values.

b. *Addition of New Mappings:*
Any new mappings derived from the output Auto Mapped Codelist are added to the Knowledge Bank, ensuring its enrichment with the latest mapping data.

a. *Resolution of Conflicting Mappings:*
In instances where a raw value exists in both the output automapped codelist and the SDTM Knowledge Bank, mapped to different CDISC Submission Values, the Update SDTM Codelist Knowledge Bank program prioritizes the new value. This ensures that the SDTM Knowledge Bank reflects the most current and accurate mappings, thus facilitating consistency and alignment with evolving data standards.

2. **Version-Specific Knowledge Banks:**
It is noteworthy that unique SDTM Codelist Knowledge Banks are stored for each different version of the SDTM Controlled Terminology. When the Codelist Auto Mapping Program is run, it checks the version of the Controlled Terminology used and selects the corresponding Knowledge Bank for optimal results. This version-specific approach ensures the accuracy and relevance of mapping data, tailored to the specific version of the SDTM Controlled Terminology being utilized.

SDTM PROGRAMMING MAPPING SPECIFICATION AND DEFINE XML:

The SDTM Programming Mapping Specification is generated by spec_creator tool, which is based on thoroughly reviewed and updated Mapping Meta Specification. **Display 8** showcases the generated SDTM Programming Mapping Specification, which includes dataset name referred to as sheet name, and all other standard structure for SDTM SAS data sets that are to be submitted to regulatory authorities. Additionally, it's also included a metadata sheet (**Display 9**) to specify different origins, which is another key requirement for FDA submission requirements (Evgeny Starostin, 2019).

The format of the final SDTM in the SDTM Programming Mapping Specification mirrors from the auto-generated codelist. The conversion definition is derived from the Mapping Meta Specification by concatenating the Formoid and Fieldoid, representing the raw dataset and variable names. The format names and origins are extracted from the Annotated ALS and external sources. In the streamlined process, updates for non-derived variables are solely required from the origin of the Annotated ALS. By rerunning the process, all updates seamlessly transfer to the codelist and mapping specifications. These updates can subsequently be integrated into final SDTM programs, ensuring a consistent and stable process. SDTM programs can be created based the variable, conversion definitions, and the automatically generated codelist as format.

EG Domain Mapping Specifications									
Protocol Number:	DEMO-XXX-XXXX								
SDTM IG Version:	3.3								
SDTM Domain Description:	ECG Test Results								
SDTM Domain Structure:	One record per ECG observation per replicate per time point or one record per ECG observation per beat per visit per								
Program Name:	eg.sas								
General Observation Class:									
Variable Name	**Variable Label**	**Type**	**Length**	**Controlled Terms or Format**	**Origin**	**Core**	**Conversion Definition**	**Variable Type**	**Variable Order**
STUDYID	Study Identifier	text	$200		Protocol	Req	DEMO-XXX-XXXX	SDTM	1
DOMAIN	Domain Abbreviation	text	$2	DOMAIN	Assigned	Req	EG	SDTM	2
USUBJID	Unique Subject Identifier	text	$200		eDT	Req	raw.CECG.USUBJID	SDTM	3
EGSEQ	Sequence Number	float			eDT	Req	raw.CECG.EGSEQ	SDTM	9
EGREFID	ECG Reference ID	text	$200		eDT	Perm	raw.CECG.EGREFID	SDTM	11
EGTESTCD	ECG Test or Examination Short Name	text	$8	EGTESTCD		Req	raw.CECG.EGTESTCD;raw.EG.EGOVRES	SDTM	17
EGTEST	ECG Test or Examination Name	text	$40	EGTEST	eDT	Req	raw.CECG.EGTEST	SDTM	18
EGCAT	Category for ECG	text	$200	EGCAT	eDT	Perm	raw.CECG.EGCAT	SDTM	27
EGPOS	ECG Position of Subject	text	$200	POSITION	eDT	Perm	raw.CECG.EGPOS	SDTM	29
EGORRES	Result or Finding in Original Units	text	$200	EGORRES		Exp	raw.CECG.EGORRES;raw.EG.EGOVRES	SDTM	31
EGORRESU	Original Units	text	$200	UNIT	eDT	Perm	raw.CECG.EGORRESU	SDTM	32
EGSTRESC	Character Result/Finding in Std Format	text	$200		eDT	Exp	raw.CECG.EGSTRESC	SDTM	40
EGSTRESN	Numeric Result/Finding in Standard Units	float			eDT	Perm	raw.CECG.EGSTRESN	SDTM	41
EGSTRESU	Standard Units	text	$200	UNIT	eDT	Perm	raw.CECG.EGSTRESU	SDTM	42
EGSTAT	Completion Status	text	$8	ND	CRF	Perm	raw.EG.EGOCCUR	SDTM	53

Display 8. Codelist Meta Specification

	A	B	C	D	E	F
1	DATASET	NAME	DSOURCE	TEST	PROGRAMMING_NOTE	
2	EG	EGDTC	CRF		raw.EG.EGDAT	
3	EG	EGDTC	External		raw.CECG.EGDTC	
4	EG	EGORRES	CRF	INTP	raw.EG.EGOVRES	
5	EG	EGORRES	External	AVCOND	raw.CECG.EGORRES	
6	EG	EGORRES	External	AXISVOLT	raw.CECG.EGORRES	
7	EG	EGORRES	External	CHYPTENL	raw.CECG.EGORRES	
8	EG	EGORRES	External	EGHRMN	raw.CECG.EGORRES	
9	EG	EGORRES	External	IVTIACD	raw.CECG.EGORRES	
10	EG	EGORRES	External	MI	raw.CECG.EGORRES	
11	EG	EGORRES	External	PACEMAKR	raw.CECG.EGORRES	
12	EG	EGORRES	External	PRAG	raw.CECG.EGORRES	
13	EG	EGORRES	External	QRSAG	raw.CECG.EGORRES	
14	EG	EGORRES	External	QTAG	raw.CECG.EGORRES	
15	EG	EGORRES	External	QTCBAG	raw.CECG.EGORRES	
16	EG	EGORRES	External	QTCFAG	raw.CECG.EGORRES	
17	EG	EGORRES	External	RRAG	raw.CECG.EGORRES	
18	EG	EGORRES	External	SNRARRY	raw.CECG.EGORRES	
19	EG	EGORRES	External	SPRARRY	raw.CECG.EGORRES	
20	EG	EGORRES	External	SPRTARRY	raw.CECG.EGORRES	
21	EG	EGORRES	External	STSTWUW	raw.CECG.EGORRES	
22	EG	EGORRES	External	TECHQUAL	raw.CECG.EGORRES	
23	EG	EGORRES	External	VTARRY	raw.CECG.EGORRES	
24	EG	EGORRES	External	INTP	raw.CECG.EGORRES	
25	EG	EGTESTCD	CRF		raw.EG.EGOVRES	
26	EG	EGTESTCD	External		raw.CECG.EGTESTCD	

Display 9. SDTM Meta Data

Once the SDTM specification, SDTM datasets, and automatically generated codelist containing submission values are prepared, we proceed to utilize the define tool to generate our define.xml. Leveraging the comprehensive details outlined in the SDTM Programming Mapping Specification, the resulting define.xml accurately represents the source and value levels. It precisely presents the codelist terms derived from the CRF dropdown list and external edt (**Display 10**).

AEACN	Action Taken with Study Treatment	text	Record Qualifier	ACN	Action Taken with Study Treatment • "DOSE NOT CHANGED" = "Dose Not Changed" • "DRUG INTERRUPTED" = "Drug Interrupted" • "DRUG WITHDRAWN" = "Drug Withdrawn" • "NOT APPLICABLE" = "Not Applicable" • "OTHER" = ""	Collected (Source: Investigator) Annotated Case Report Form [81]

EGTESTCD VLM		ECG Test or Examination Short Name	text	Topic	EGTESTCD	ECG Test Code [20 Terms]	
	EGGRPID = "LOCAL"	Local	text		EGTESTCD	ECG Test Code [20 Terms]	Collected (Source: Investigator) Annotated Case Report Form [52]
	EGGRPID = "EXTERNAL"	External	text		EGTESTCD	ECG Test Code [20 Terms]	Collected (Source: Vendor)
EGORRES VLM		Result or Finding in Original Units	text	Result Qualifier	45		
	EGTESTCD = "AVCOND"	Atrioventricular Conduction	text		21		Collected (Source: Vendor)
	EGTESTCD = "AXISVOLT"	Axis and Voltage	text		11		Collected (Source: Vendor)
	EGTESTCD = "CHYPTENL"	Chamber Hypertrophy or Enlargement	text		28		Collected (Source: Vendor)
	EGTESTCD = "EGHRMN"	ECG Mean Heart Rate	integer		8		Collected (Source: Vendor)
	EGTESTCD = "INTP" and EGGRPID = "LOCAL"	Interpretation	text		38	Interpretation • "ABNORMAL" • "NORMAL"	Collected (Source: Investigator) Annotated Case Report Form [52]

Display 10. Final Define xml

COLUSION:

BENEFITS:

The streamlined data driven ALS to SDTM process has yielded significant improvements in efficiency, accuracy, and consistency. By reducing manual intervention and leveraging automated algorithms, the process expedites mapping tasks, mitigates errors, and ensures compliance with CDISC standards. Furthermore, the systematic approach enhances reproducibility and scalability across diverse datasets and studies.

An added benefit of this process is its iterative improvement over time. With each utilization during studies, the program's efficacy grows. As the SDTM ALS/Codelist Knowledge Bank accumulates more prior mappings, the program's accuracy proportionally increases, ultimately reducing the need for manual intervention.

FUTURE WORK:

Currently under development, this program is poised to introduce additional features aimed at further enhancing its functionality and adaptability to diverse mapping needs. One such feature is the capability for programmers to select certain past studies as preferential mapping references. Recognizing that different studies may entail unique mapping requirements, this hierarchical approach to the Knowledge Bank mapping will enable more precise and tailored mapping solutions. By empowering users to prioritize specific studies based on relevance and similarity to the current project, this enhancement promises to

refine mapping accuracy and streamline the automation process further. Continuous development efforts are dedicated to bringing this feature to fruition, thereby ensuring the SDTM Mapping Automation framework's sustained advancement and optimization.

REFERENCES

Evgeny, Starostin. 2019. “Quality Control and Validation – More than Just PROC COMPARE.” <*Phuse Connect*>, <10th-13th no.:7>.

ACKNOWLEDGMENTS

I extend my sincere gratitude and acknowledgment to the information technology team and programming team at ClinChoice for their unwavering dedication and hard work in developing tools alongside ongoing study tasks. Additionally, special thanks to the Regeneron Team for providing feedback and insight throughout the development of these automation tools.

CONTACT INFORMATION

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

Yunsheng Wang
ClinChoice Inc.
1300 Virginia Dr,
Fort Washington, PA 19034
E-mail: yunsheng.wang@clinchoice.com
Web : https://clinchoice.com/