

CDISC SDTM CONVERSION IN ISS/ISE STUDIES: TOOLS

Balaji Ayyappan, PharmaNet/i3, Cary, NC

ABSTRACT:

CDISC SDTM (Study Data Tabulation Model) conversion for Integration Summary of Safety (ISS) and Integrated Summary of Efficacy (ISE) is always a challenging job. The most taxing task is the standardization of different non-CDISC compliant raw datasets to CDISC compliant domains and variables for all the legacy studies involving large amount data and diverse systems of data collection. Having faced this challenge year after year, I have been involved in the development of a set of tools and reports to perform and validate the SDTM conversion for ISS/ISE studies of varying proportion to complete the task in a manner that is efficient and accurate. This paper will describe the functionalities of this utility. The technologies used are SAS, Microsoft Excel, VBA and SAS IOM. Knowledge of VBA and SAS IOM fundamentals is a plus, but not a limiting factor.

INTRODUCTION:

ISS and ISE analyses are considered as some of the most important components for the FDA NDA submission. When we have tools to check SDTM process implementation at various stages, it makes the process easier and time efficient. This paper will explain the tools and reports generated at various stages of ISS/ISE implementation:

- i. Stage I : Pre-processing
- ii. Stage II : SDTM Implementation
- iii. Stage III: Post-Processing

STAGE I (PRE-PROCESSING):

In this stage, we will be scrutinizing different scenarios in order to achieve accurate integrated datasets for the ISS/ISE analysis. In the initial stage of the implementation it will be very useful to group the studies with similar data structures.

In order to find similar studies or common data structures across the studies, we have developed a macro which takes the metadata information (ie. raw dataset name, variable name, variable attributes like label and format) and checks it to derive the percentage of structural homogeneity. The higher the percentage (number within the parentheses following the dataset name), the higher the similarity. A zero would indicate there is no matching dataset.

For example, if we have a raw AE dataset in study 1, this dataset's Meta information is compared with all other raw AE datasets. Then another AE dataset is selected and compared against the rest, and so on. This process is followed for all the datasets. At the end, we will have a report with the percentage of similarity for each raw dataset across the studies in an nXn grid, where n is the number of studies. The screen shot below shows a sample report. Percentages are displayed in parenthesis along with each raw dataset.

Study	Study 1	Study 2	Study 3	Study 4
Study 1	AE(52), CHEM(0), CHEMII(0), CMPCODED(23.8), CONILL(0), CONMED(11.8), DEMOG(57.9), DRUGTST(0), DSTOOL(0), DTRT(0), EFF2(0), EFFICACY(0), ENTERED(66.7), ENTRY(30.4), EVAL(0), GISYMP(38.9), HEMAT(0), MEDHX(9.3), MEDREC(0), PCTCHG(0), PE(45.3), RANGES(0), STCHLOR(0), STCLCONC(0), STOOLCHL(0), TERM(42.9), TRT(0), URIN(0)	AE(88), CHEM(0), CHEMII(0), CMPCODED(0), CONILL(81.3), CONMED(82.4), DEMOG(0), DRUGTST(0), DSTOOL(0), DTRT(0), EFF2(0), EFFICACY(0), ENTERED(100), ENTRY(47.8), EVAL(0), GISYMP(66.7), HEMAT(0), MEDHX(5.6), MEDREC(13.7), PCTCHG(0), PE(18.9), RANGES(0), STCHLOR(0), STCLCONC(0), STOOL(65.4), STOOLCHL(0), TERM(85.7), TRT(33.3), URIN(0)	AE(84), CHEM(0), CHEMII(0), CMPCODED(61.9), CONILL(0), CONMED(52.9), DEMOG(0), DRUGTST(0), DSTOOL(0), DTRT(0), EFF2(0), EFFICACY(0), ENTERED(100), ENTRY(47.8), EVAL(0), GISYMP(66.7), HEMAT(0), MEDHX(5.6), MEDREC(13.7), PCTCHG(0), PE(28.3), RANGES(0), STCHLOR(0), STCLCONC(0), STOOL(65.4), STOOLCHL(0), TERM(78.6), TRT(33.3), URIN(0)	
Study 2	CMPCODED(15.2), CMPCODES(0), CONMED(6.3), DAY1435(0), DAY28(0), DAY7(0), DEMOG(56.4), DTRT(0), ENTERED(57.1), ENTRY(41.2), ENTRY2(0), EVAL(0), GISYMP(13), HEMAT(0), LABSTOOL(0), MED(0), MEDHX(20.8), OTHERLAB(0), PATIENTS(0), PE(42.9), PKTRT(0), PKURINE(0), REGSTAT1(0), REGSTAT2(0), RESPOND(0), RTRT(0), STLCNTS(0), STLOUT(0), STOOL(3), TERM(28.6), TRT(0), URIN(0), VITALS(0)		CMPCODED(0), CMPCODES(0), CONMED(6.3), DAY1435(0), DAY28(0), DAY7(0), DEMOG(0), DTRT(0), ENTERED(57.1), ENTRY(41.2), ENTRY2(0), EVAL(0), GISYMP(13), HEMAT(0), LABSTOOL(0), MED(0), MEDHX(8.3), OTHERLAB(0), PATIENTS(0), PE(17.9), PKTRT(0), PKURINE(0), REGSTAT1(0), REGSTAT2(0), RESPOND(0), RTRT(0), STLCNTS(0), STLOUT(0), STOOL(3), TERM(28.6), TRT(0), URIN(0), VITALS(0)	AE(52), BLSTLWT(0), CHEM(0), CMPCODED(21.2), CMPCODES(0), CONMED(9.4), DAY1435(0), DAY28(0), DAY7(0), DEMOG(0), DTRT(0), ENTERED(57.1), ENTRY(41.2), ENTRY2(0), EVAL(0), GISYMP(13), HEMAT(0), LABSTOOL(0), MED(0), MEDHX(8.3), OTHERLAB(0), PATIENTS(0), PE(25), PKTRT(0), PKURINE(0), REGSTAT1(0), REGSTAT2(0), STLCNTS(0), RESPOND(0), RTRT(0), STLCNTS(0), STLOUT(0), STOOL(3), TERM(28.6), TRT(0), URIN(0), VITALS(0)
Study 3	AE(88), BLEVAL(0), CONILL(81.3), CONMED(82.4), ENTERED(100), ENTRY(50), GISYMP(60), MEDHX(13), MEDREC(62.5), PDR(0), PDRSS(0), PE(58.8), RANDOM(0), STOOL(63), TERM(60), TRT(25)	AE(52), BLEVAL(0), CONILL(0), CONMED(11.8), ENTERED(66.7), ENTRY(31.8), GISYMP(35), MEDHX(8.7), MEDREC(0), PDR(0), PDRSS(0), PE(58.8), RANDOM(0), STOOL(14.8), TERM(30), TRT(0)		AE(96), BLEVAL(81.5), CONILL(0), CONMED(41.2), ENTERED(100), ENTRY(100), GISYMP(95), MEDHX(84.8), MEDREC(87.5), PDR(80.9), PDRSS(68), PE(70.6), RANDOM(0), STOOL(100), TERM(95), TRT(100)
Study 4	AE(80.8), BLEVAL(0), CMPCODED(61.9), CONMED(52.9), DUMTRT(0), ENTERED(100), ENTRY(50), GISYMP(57.1), MEDHX(9.5), MEDREC(58.8), PDR(0), PDRSS(0), PE(65.2), STOOL(63), TERM(52.4), TRT(25), TRT2(0)	AE(50), BLEVAL(0), CMPCODED(33.3), CONMED(17.6), DUMTRT(0), ENTERED(66.7), ENTRY(31.8), GISYMP(33.3), MEDHX(6.3), MEDREC(0), PDR(0), PDRSS(0), PE(60.9), STOOL(14.8), TERM(28.6), TRT(0), TRT2(0)	AE(92.3), BLEVAL(62.9), CMPCODED(0), CONMED(41.2), DUMTRT(0), ENTERED(100), ENTRY(100), GISYMP(90.5), MEDHX(61.9), MEDREC(82.4), PDR(76), PDRSS(85), PE(52.2), STOOL(100), TERM(90.5), TRT(100), TRT2(0)	

Before programming, key information about the study design is pooled together in a spread sheet, like study drug, treatment period, frequency dosing information and total dose amount per day.

We can also add ARMCD (planned treatment) and pool column. This sheet will help us provide a quick understanding of each study to the programmers/statistician involved in the analyses and also can be used for Trial Design datasets. This sheet is created manually by going through the protocol/SAP for each study in the analyses.

Please check the screen shot below which gives the key information about the design for all the studies.

Study No	Study Drug	Description							ARMCD	
		Treatment Period	Dose Timings		Dose Taken/Interval			Total Dose Amount Per Day		
			Per Day	Total	Dose Amount in Single Capsule	No. of capsules/ Tablets	Dose Amount			
Study 1	ABC Delayed Release Beads	4 days (96 Hours)	4 times (Every 6 hrs.)	16 Consecutive Doses	250 mg	2	500 mg	2000 mg	ABC	
	PLACEBO	4 days (96 Hours)	4 times (Every 6 hrs.)	16 Consecutive Doses	0 mg	2	0 mg	0 mg	PBO	
Study 2	ABC Enteric Coated Tablets	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	50 mg	1	50 mg	200 mg	ABC1	
	ABC Enteric Coated Tablets	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	50 mg	3	150 mg	600 mg	ABC2	
	PLACEBO	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	0 mg	1	0 mg	0 mg	PBO1	
	PLACEBO	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	0 mg	3	0 mg	0 mg	PBO2	
Study 3	250 mg ABC Enteric coated Beads	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	250 mg (enteric coated beads)	1 Capsule	250 mg	1000 mg	ABC1	
	250 mg ABC Tablet	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	250 mg (2*125 mg ABC Tablets)	1 Capsule	250 mg	1000 mg	ABC2	
	500 mg ABC Tablet	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	500 mg (4*125 mg ABC Tablets)	1 Capsule	500 mg	2000 mg	ABC3	
	PLACEBO	2 Days (48 Hours)	4 times (Every 6 hrs.)	8 Consecutive Dose	0 mg	1 Capsule	0 mg	0 mg	PBO	

When we have a good understanding about the design of each study and knowledge about similar studies, it will make it easier for the team to accurately integrate the datasets for the ISS/ISE analyses. It is advisable to collect all the unique raw dataset values for key variables to which control terminologies are to be applied. For example, collect all the unique DSTERM values in the spreadsheet for which control terminology is applied. DSDECOD is derived. The same is true for AEACN, AEOUT, LBTEST, etc.. The sheet can then be used for creating the SAS formats while implementing SDTM conversion across the study.

STAGE II (SDTM IMPLEMENTATION):

During the programming stage for SDTM conversion across the studies, the tool is used to create a CDISC Variable mapping and a Control Terminology document. This describes, at study level, how the CDISC variables are mapped to the raw dataset variable(s) for each domain across the studies. Standardizing the unique values for certain variables in standard CDISC domains is documented in Control Terminology (CT).

The ISS CDISC variable mapping document will have sheets to contain the Project Directory paths (Directories like: Raw Dataset, CDISC Dataset, SAS Program, etc.) and mapping information of Raw dataset(s) for each domain. Sheets for each Domain, which are to be mapped, will be present in this spreadsheet. There are pre-defined columns, like CDISC variable name, label, type, length, CT format, if applicable. For more details about CDISC variable mapping and CT document tool, please refer to "CDISC Variable Mapping and Control Terminology Implementation Made Easy" (PharmaSUG2011 - Paper CD11, <http://www.lexjansen.com/pharmasug/2011/cd/pharmasug-2011-cd11.pdf>)

The main advantage of this document is that we know how each raw variable is mapped to a CDISC variable for the corresponding domain across the studies. It helps us to avoid mistakes while mapping the raw dataset variables. Please check the screen shot of the variable mapping document for ISS studies below.

VARNAM	VARLBL	TYPE	LEN	FMT	ORIGIN	ROLE	7447_DATAMAP	7447_COMMENTS	7526_DATAMAP	7526_COMMENTS	7527_DATAMAP
STUDYID	Study Identifier	Char	\$6			Identifier		"7447-U-01"		"7526-U-01"	
DOMAIN	Domain Abbreviation	Char	\$2	DM		Identifier		"DM"		"DM"	
USUBJID	Unique Subject Identifier	Char	\$16			Identifier	DEMOGRAPHIC.CENTER(\$1), DEMOGRAPHIC.K_PATID()		DEMOGRAPHIC.CENTER(\$1), DEMOGRAPHIC.K_PATID()		DEMOGRAPHIC.K_PATID(), DEMOGRAPHIC.CENTER(\$18)
SUBJID	Subject Identifier for the Study	Char	\$5			Topic	DEMOGRAPHICK_PATID()		DEMOGRAPHICK_PATID()		DEMOGRAPHICK_PATID()
RFSTDTDC	Subject Reference Start Date/Time	Char	\$19	ISO 8601		Record Qualifier	VS.DATA(DATE3), VS.CHKPT()		VS.Visit(), VS.Date_Of_Visit(DATE3)		VS.CHKPT(), VS.DATA(DATE3)
RFENDTC	Subject Reference End Date/Time	Char	\$19	ISO 8601		Record Qualifier	VS.DATA(DATE3), VS.CHKPT()		END.TERMDATE(DATE3), VS.Visit(), VS.Date_Of_Visit(DATE3)		END.TERMDATE(DATE3), VS.CHKPT(), VS.DATA(DATE3)
SITEID	Study Site Identifier	Char	\$10			Record Qualifier	DEMOGRAPHIC.CENTER(\$15)		DEMOGRAPHIC.CENTER(\$1)		DEMOGRAPHIC.CENTER(\$18)
INVID	Investigator Identifier	Char	\$15			Record Qualifier					
INVNAM	Investigator Name	Char	\$30			Synonym Qualifier					
BRTHDTDC	Date/Time of Birth	Char	\$19	ISO 8601		Record Qualifier	DEMOGRAPHIC.BIRTHDATE(DATE3)				DEMOGRAPHIC.BIRTHDATE(DATE3)
AGE	Age	Num	8			Record Qualifier	DEMOGRAPHIC.AGE()		DEMOGRAPHIC.AGE()		DEMOGRAPHIC.BIRTHDATE(DATE3)
AGEU	Age Units	Char	\$10	(AGEU)		Variable Qualifier	RANDOM.CTD_code()	"YEARS"		"YEARS"	
SEX	Sex	Char	\$1	(SEX)		Record Qualifier	RANDOM.CTD_code(), RANDOM.Patient(), RANDOM.Treatment(\$10), VS.STUDY(\$4), VS.PHASE(\$3), VS.K_PATID(), VS.CENTER(\$15)		DEMOGRAPHIC.SEX()		DEMOGRAPHIC.SEX()
RACE	Race	Char	\$50			Record Qualifier			DEMOGRAPHIC.RACE()		DEMOGRAPHIC.RACE()

I \ \ SETUP \ INDEX \ DM \ CM \ EX \ AE \ DS \ EG \ LB \ SC \ SUPPQUAL \ VS \ MB \ MS \ XF \ XS \ XD \ XM \ CT \ Sheet1 \ CT \ [] []

A	B	C	D	E	F	G	H	I	J
Study	Domain	CDISC Format Name	CDISC Variable Name	Raw Variable Name (Label)	Raw Variable	Raw Variable Informat Values	Raw Variable Format Values	Controlled Terminology	Comment
N105	AE	(NY)	AE SER	AESAES.SAEPREF (AE: SAE - present-FUL)	char	NO	No	N	
N105	AE	(NY)	AE SER	AESAES.SAEPREF (AE: SAE - present-FUL)	char	YES	Yes	Y	
N105	AE	(AESEV)	AE SEV	AESAES.AESEV	char	MIL	Mild	MILD	
N105	AE	(AESEV)	AE SEV	AESAES.AESEV	char	MOD	Moderate	MODERATE	
N105	AE	(AESEV)	AE SEV	AESAES.AESEV	char	SEV	Severe	SEVERE	

I \ \ SETUP \ INDEX \ DM \ CM \ EX \ AE \ DS \ EG \ LB \ SC \ SUPPQUAL \ VS \ MB \ MS \ XF \ XS \ XD \ XM \ Sheet1 \ CT \ [] []

STAGE III(POST-PROCESSING):

After completing SDTM conversions for all the studies, it is a good practice to do some post-processing checks to make sure integrated datasets are accurate. The report created will show unique SDTM dataset values with respect to the individual study (a.k.a "Study Wise") and also for integrated datasets. For example, if we want to check the values of LBCAT, LBTEST & LBTESTCD when we pass these variables in the macro, we will have the following report.

S. No.	Study 1(LBCAT)	Study 1(LBTEST)	Study 1(LBTESTCD)	Study 2(LBCAT)	Study 2(LBTEST)	Study 2(LBTESTCD)
1	CHEMISTRY	Alanine Aminotransferase	ALT	CHEMISTRY	Alanine Aminotransferase	ALT
2	CHEMISTRY	Albumin	ALB	CHEMISTRY	Albumin	ALB
3	CHEMISTRY	Alkaline Phosphatase	ALP	CHEMISTRY	Alkaline Phosphatase	ALP
4	CHEMISTRY	Amylase	AMYLASE	CHEMISTRY	Amylase	AMYLASE
5	CHEMISTRY	Aspartate Aminotransferase	AST	CHEMISTRY	Aspartate Aminotransferase	AST
6	CHEMISTRY	Bilirubin	BILI	CHEMISTRY	Bicarbonate	BICARB
7	CHEMISTRY	Blood Urea Nitrogen	BUN	CHEMISTRY	Bilirubin	BILI
8	CHEMISTRY	Calcium	CA	CHEMISTRY	Blood Urea Nitrogen	BUN
9	CHEMISTRY	Carbon Dioxide	CO2	CHEMISTRY	Calcium	CA
10	CHEMISTRY	Chloride	CL	CHEMISTRY	Chloride	CL
11	CHEMISTRY	Creatinine	CREAT	CHEMISTRY	Creatinine	CREAT
12	CHEMISTRY	Creatinine Clearance	CREATCLR	CHEMISTRY	Creatinine Clearance	CREATCLR
13	CHEMISTRY	Direct Bilirubin	BILDIR	CHEMISTRY	Direct Bilirubin	BILDIR
14	CHEMISTRY	Glucose	GLUC	CHEMISTRY	Glucose	GLUC
15	CHEMISTRY	Indirect Bilirubin	BILIND	CHEMISTRY	Iron	IRON
16	CHEMISTRY	Iron	IRON	CHEMISTRY	Lactate Dehydrogenase	LDH
17	CHEMISTRY	Lactate Dehydrogenase	LDH	CHEMISTRY	Magnesium	MG
18	CHEMISTRY	Magnesium	MG	CHEMISTRY	Phosphate	PHOS
19	CHEMISTRY	Phosphate	PHOS	CHEMISTRY	Potassium	K
20	CHEMISTRY	Potassium	K	CHEMISTRY	Protein	PROT
21	CHEMISTRY	Protein	PROT	CHEMISTRY	Sodium	SODIUM
22	CHEMISTRY	Sodium	SODIUM	DRUG SCREEN	Amphetamine	AMPHET
23	CHEMISTRY	Urate	URATE	DRUG SCREEN	Barbiturate	BARB
24	HEMATOLOGY	Basophils/Leukocytes	BASOLE	DRUG SCREEN	Benzodiazepine	BNZDZPN
25	HEMATOLOGY	Eosinophils/Leukocytes	EOSLE	DRUG SCREEN	Cannabinoids	CANNAB
26	HEMATOLOGY	Erythrocytes	RBC	DRUG SCREEN	Cocaine	COCAINE
27	HEMATOLOGY	Hematocrit	HCT	DRUG SCREEN	Ethanol	ETHANOL
28	HEMATOLOGY	Hemoglobin	HGB	DRUG SCREEN	Methadone	METHDN
29	HEMATOLOGY	Leukocytes	WBC	DRUG SCREEN	Opiate	OPIATE
30	HEMATOLOGY	Lymphocytes	LYM	DRUG SCREEN	Phencyclidine	PCP
31	HEMATOLOGY	Lymphocytes Atypical/Leukocytes	LYMATLE	HEMATOLOGY	Activated Partial Thromboplastin Time	APTT
32	HEMATOLOGY	Lymphocytes/Leukocytes	LYMLE	HEMATOLOGY	Basophils	BASO
33	HEMATOLOGY	Monocytes/Leukocytes	MONOLE	HEMATOLOGY	Basophils/Leukocytes	BASOLE
34	HEMATOLOGY	Neutrophils	NEUT	HEMATOLOGY	Blasts	BLAST
35	HEMATOLOGY	Neutrophils Band Form/Leukocytes	NEUTBLE	HEMATOLOGY	Blasts/Leukocytes	BLASTLE
36	HEMATOLOGY	Neutrophils , Segmented/Leukocytes	NEUTSGLE	HEMATOLOGY	Eosinophils	EOS
37	HEMATOLOGY	Platelet	PLAT	HEMATOLOGY	Eosinophils/Leukocytes	EOSLE
38	HEMATOLOGY	Prothrombin Intl. Normalized Ratio	INR	HEMATOLOGY	Erythrocytes	RBC
39	HEMATOLOGY	Prothrombin Time	PT	HEMATOLOGY	HAV PCR Viral Load	HAVPCR
40	HEMATOLOGY	Thrombin Time	TT	HEMATOLOGY	HEP-A VIRAL AB.(IGM)	HEP3
41	HIV	CD4	CD4	HEMATOLOGY	HEP-B CORE AB IGM	HEP4
42	HIV	CD8	CD8	HEMATOLOGY	HEP-B CORE ANTIBODY	HEP5
43	HIV	HIV PCR Viral Load	HIVPCR	HEMATOLOGY	Hematocrit	HCT
44	URINALYSIS	Bile Acid	BILEAC	HEMATOLOGY	Hemoglobin	HGB
45	URINALYSIS	Choriongonadotropin Beta	HCG	HEMATOLOGY	Hepatitis B Virus Surface Antigen	HBSAG

The Study Wise sheet in the report will show us the unique values of LBCAT, LBTEST, and LBTESTCD and ensures control terminology is applied appropriately across the studies. The Unique Value sheet will provide information for the integrated ISS dataset. This post-process will allow us to check and make sure all variables are standardized and compliant.

CONCLUSION:

These different tools are used at various stages of implementation to achieve the accuracy of the integrated datasets in a time efficient manner. It is useful for creating CDISC variable and CT mappings with no typing errors. These tools will drastically reduce the time for preparing documentation. These documents can also be used as a source for creating the DEFINE.XML file.

REFERENCES:

Using VBA and BASE SAS to Get Data from SAS to Excel without Data Integrity Issues
<http://www.lexjansen.com/phuse/2005/as/as11.pdf>

CDISC Variable Mapping and Control Terminology Implementation made easy
<http://www.lexjansen.com/pharmasug/2011/cd/pharmasug-2011-cd11.pdf>

ACKNOWLEDGEMENTS:

I would like to thank Paul Ragland and Shridhar Patel for providing excellent ideas and helping to implement this paper.

CONTACT INFORMATION:

Your comments and questions are valued and encouraged. You can contact at:

Balaji Ayyappan
Phrmanet/i3,
1001 Winstead Drive,
Cary, NC 27513
Phone: (919) 678 4705
Email: bayyappan@pharmanet-i3.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.