Outsourced Data Integration Project with CDISC SDTM & ADaM Deliverables

Christine Teng, Merck Research Labs, Merck Sharp & Dohme Corp., Rahway, NJ
Margaret Coughlin, Merck Research Labs, Merck Sharp & Dohme Corp., Rahway, NJ

ABSTRACT

The Clinical Data Interchange Standards Consortium (CDISC) has established platform-independent data standards that enable information systems interoperability to improve clinical research and related areas of healthcare. Many pharmaceutical companies have started implementing CDISC clinical trial data models such as the Study Data Tabulation Model (SDTM) and the Analysis Data Model (ADaM). The purpose of this paper is to discuss some experience gained from working with a CRO on a data integration project that converts several studies into a common SDTM structure in support of eCTD and Integrated Summary of Safety analyses. As this is a relatively early project working with a CRO on SDTM integration, processes of working with CROs in this area are still evolving. This paper briefly describes important components that are recommended to be provided to a CRO in order to facilitate the process of mapping data to the CDISC data models. The paper will also discuss review activities to help verify that CRO-converted SDTM datasets comply with CDISC standards.

SAS®9, Windows®, Intermediate Level
Key Words: CDISC, SDTM, ADaM

BACKGROUND

The data integration project was comprised of several studies that needed integrated safety analysis for a filing. These studies were not done internally so eCRFs were not designed uniformly nor was it likely that the ultimate goal of converting the data to SDTM was ever considered. The availability of standard submission data in SDTM format provides a solution for the purpose of integration that also benefits regulatory reviewers, as the FDA has put in considerable effort to develop a repository for all submitted trial data and a suite of standard review tools to access, manipulate, and view the tabulations.

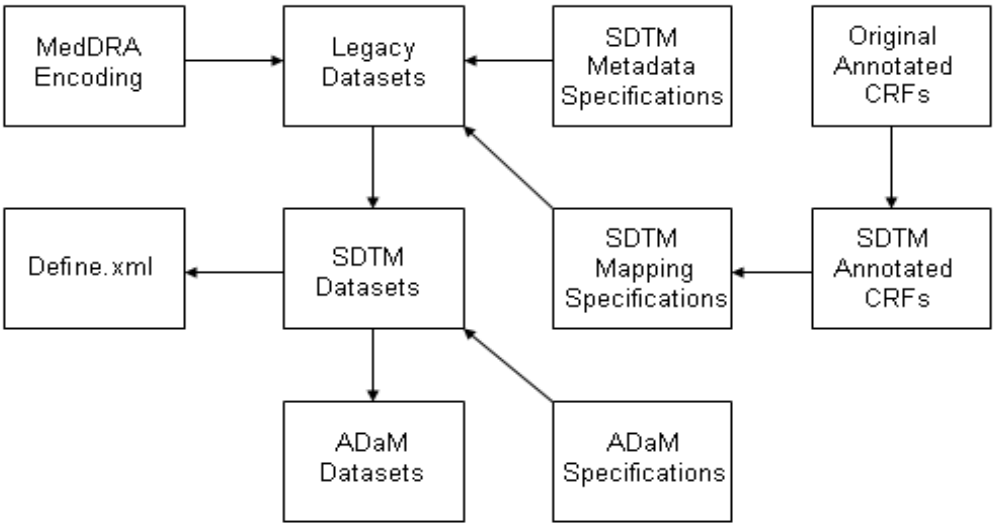
SETTING EXPECTATIONS

Accurate mapping of study data to SDTM format is critical for a successful SDTM conversion and integration process. In order to facilitate a smooth mapping process, it is helpful that some basic expectations are established with the CRO. The expectations provided by the sponsor should provide a clear picture of what a successful SDTM conversion and integration will look like and assist the partner in working toward that goal.

Each company has its own standards and SOPs. It is helpful to provide specific basic expectations to ensure that the deliverables (output datasets, tables/listings/figures, and programs) can be reproduced internally if the intention is to reuse the programs developed by the CRO at a later time. Situations to consider when deciding if internal reproducibility is necessary include whether the sponsor is responsible for responses to agency questions, publications, or other external presentations. For example, to minimize modifications to the CRO-supplied programs, it is helpful that no hard coded library or file path is defined inside the programs. Also, configurations should be done in one autoexec-like program so that changes can be done in one location instead of in many programs. This is important if there is a need to reuse the CRO-supplied programs to verify the deliverables.

To help ensure consistency, data handling or mapping rules should be defined up front. Examples include how to handle missing data and numeric character test results, such as >3000 or 25mg. Other examples include keeping only needed variables in analysis datasets and setting efficient lengths for character data when the default length greatly exceeds the longest value. We recommend that a well-prepared package of deliverables consist of validated programs, efficient code, consistent code and reproducible code.

The high level process map to create the deliverables is depicted as follows:



- Beginning in the upper right-hand box, the original annotated CRF referenced to the legacy data are converted to reference SDTM domains and variables. The SDTM annotated CRFs are created for this purpose.
- Next, the SDTM mapping specifications, in xls format, are created based on the SDTM annotated CRFs.
- The SDTM metadata specifications are define.xls-like documents that are used by sponsor internally.
- The conversion programs, based on information created above and the latest MedDRA Encoding, are used to build the SDTM datasets from the legacy data.
- For submission, the define.xml is generated by the validation tool from the SDTM datasets.
- ADaM specifications are built from the SDTM datasets for integrated analysis purpose.
- ADaM datasets are created using ADaM specifications.

Further definitions of the items presented here are found in the next section**.**

SUPPORTING DOCUMENTATION

Both the sponsor and the CRO need to work closely and collaboratively to have a successful filing. Adequate information should be provided to the CRO as early as possible to allow time to clarify questions. If data integration is the focus, then the standardization of common items should be noted up front such as Treatment Names, Treatment Codes, Visit Names, Visit Numbers, MedDRA Encoding Version, and coding for Disposition Status. Using the SDTM structure has helped standardize the data structure. The table below further describes some materials that are helpful for the CRO to implement the SDTM conversion.

Document or Material	Helpful Content	Comment
SDTM Implementation Guide	The guidance and version to follow	
Final Protocol	Analysis end points and study design	
Data Entry Guideline	CRF visit coding, data entry conditions, page relationships and how various entries are encoded	
Annotated CRF of Legacy data	References the SDTM variables with legacy variables	Used in mapping specification
SAS formats	Useful where studies kept a numeric value in the raw data and created a SAS format to display the text result.	Text description should be mapped rather than value.
Unit of measurements and CDISC Controlled Terminology	Common lab units needed for analysis. Other controlled terminology such as Age Unit and Race are standardized.	

SDTM Mapping Specifications Template	Pre-specified format which cross references the legacy datasets and variables and the mapped SDTM domains and variables	
SDTM Metadata Specifications Template	Pre-specified metadata which documents SDTM domains, variables, attributes, keys, code lists, and code names expected for a CDISC submission.	SDTM variable length and attributes should be consistent among all studies
ADaM Specifications Template	Pre-specified template that provides derivation logic for integrated analysis datasets	Based on SAP
Trial Design Specifications	Defines trial design domains	SE,SV,TE,TV,TI, TA
Specifications for External Transfer data	Vendor specifications for external data	laboratory results, biomarker, PK data

The above items are some materials the sponsor and CRO found helpful to share, when available, in order to facilitate SDTM conversions of data in non-SDTM formats.

REVIEW FOR SDTM CONVERSION

Informal reviews of conversion were done during development to ensure the CRO did not deviate from the requirements and standards. Certain verifications are recommended to be performed for SDTM deliverables. The high level review process is done in two steps. Please note that the review process described below is for illustration and is only done for this specific project since company standards were evolving at the time. Only some of the key review points are described below.

1) Mapping verification

A SAS program is used to verify that all CRF raw datasets and all collected variables were mapped. An Excel workbook is created by loading the mapping specifications into the SAS program to facilitate the review process by matching variables in the specification with the real datasets (See Table 1 below).

Table 1
Cross-referenced of data metadata with mapping specification

Dataset	Varname	INdict	TypeLen	CRF	crf_col_name	INmapp	label	format	Comments
		0		LAB	COLDTM	1			Need to remove from mapping spec
LAB	RECEIVED_DATE	1	Num8	LAB	RECEIVED_DATE	1	Received Date	DATE7.	
LAB	RECEIVED_TIME	1	Char8	LAB	RECEIVED_TIME	1	Received Time	$8.	
LAB	REPORT_DATE	1	Num8	LAB	REPORT_DATE	1	Report Date	DATE7.	
LAB	REPORT_TIME	1	Char8	LAB	REPORT_TIME	1	Report Time	$8.	
LAB	SI_HIGH_RANGE	1	Char10	LAB	SI_HIGH_RANGE	1	SI High Range	$10.	
LAB	SI_LOW_RANGE	1	Char10	LAB	SI_LOW_RANGE	1	SI Low Range	$10.	
LAB	SI_NON_NUMERIC	1	Char500	LAB	SI_NON_NUMERI	1	SI Non Numeric Range	$500.	
LAB	SI_RESULT	1	Char14	LAB	SI_RESULT	1	SI Result	$14.	
LAB	SI_UNIT	1	Char14	LAB	SI_UNIT	1	SI Unit	$14.	
LAB	SUBJNO	1	Char20	LAB	SUBJNO	1	Subject Number	$20.	
LAB	TEST_CLASS	1	Char1	LAB	TEST_CLASS	1	Test Class (For Lab Use)	$1.	
LAB	TEST_CODE	1	Char8	LAB	TEST_CODE	1	Test Code	$8.	
LAB	TEST_NAME	1	Char30	LAB	TEST_NAME	1	Test Name	$30.	
LAB	TEST_SPECIFIC_C	1	Char500	LAB	TEST_SPECIFIC_	1	Test Specific Comments	$500.	
LAB	TEST_SPEC_COM	1	Char35	LAB	TEST_SPEC_CO	0	Test Specific Comment Codes	$35.	ok not to map
LAB	UNSCHEDULED	1	Num8	LAB	UNSCHEDULED	1	Unscheduled	4.	
LAB	US_HIGH_RANGE	1	Char10	LAB	US_HIGH_RANGE	1	US High Range	$10.	
LAB	US_LOW_RANGE	1	Char10	LAB	US_LOW_RANGE	1	US Low Range	$10.	
LAB	US_NON_NUMERI	1	Char500	LAB	US_NON_NUMER	1	US Non Numeric Range	$500.	
LAB	US_RESULT	1	Char14	LAB	US_RESULT	0	US (Conventional) Result	$14.	Should be mapped
LAB	US_UNIT	1	Char14	LAB	US_UNIT	0	US Unit	$14.	
LAB	VISIT	1	Num8	LAB	VISIT	1	Visit	5.1	

Variables inside red boxes above are collected from real datasets via the dictionary tables. 'CRF' and 'crf_col_name' columns are from the mapping spec. 'Indict' and 'INmapp' flags are used to denote their existence in dictionary or in mapping spec. By reviewing the report and comparing the flags, the sponsor notes the comments and sends back to CRO for corrections.

The SDTM annotated CRF is also reviewed by the sponsor's internal standards team to confirm agreement of mapping domains and SUPPQUAL variables of CDISC standards and practices. For example, verifications include - whether topic variables are mapped, correct class of domain is used and whether each suppqual variable has a parent record etc.

2) SDTM conversion verification

After conversion is implemented based on the specification, the SDTM datasets are reviewed. In addition to verifying the legacy data mapping to SDTM for a random sample of subjects base on pre-specified test cases, another diagnostic program is written to display certain key variables for overall verification of mapping content. Table 2 to 4 below are the generated output (data are mock-up).

Table – 2

Metadata for each domain is displayed in each tab

A	B	C	D
Domain	**Created on**	**Observations**	**Domain Label**
CF	21MAY09:13:00:44	5073	Clinical Findings

Pos	VarName	TypeLen	Label
1	STUDYID	Char18	Study Identifier
2	DOMAIN	Char2	Domain Abbreviation
3	USUBJID	Char18	Unique Subject Identifier
4	CFSEQ	Num8	Sequence Number
5	CFGRPID	Char8	Group ID
6	CFSPID	Char8	Sponsor-Defined Identifier
7	CFTESTCD	Char8	Test Short Name
8	CFTEST	Char40	Test Name
9	CFCAT	Char100	Category for Clinical Findings
10	CFORRES	Char100	Result or Finding in Original Units
11	CFSTRESC	Char100	Character Result/Finding in Std Format
12	CFBLFL	Char1	Baseline Flag
13	VISIT	Char20	Visit Name
14	VISITNUM	Num8	Visit Number
15	CFDTC	Char19	Date/Time of Collection
16	CFDY	Num8	Study Day of Finding
17	CFEVLINT	Char19	Evaluation Interval

GRPID	Topic TESTCD Code	Test Description
ADVERSE	OCCUR	Occurrence
INVESTAS	JVPRESS	Jugular venous pressure (JVP)
INVESTAS	PEREDEMA	Peripheral edema
INVESTAS	RALES	Rales
SDADMIN	SEIZURE	Risk Factor for Seizure

AE | CF | CM | CO | DM | DS | EG | EX | HO | IE | LB | MH | PR | QS | RELREC | SC | SE | SV

Topic value for the domain is listed.

Table – 3

All TESTCDs are displayed including the Original, Standard, and Conventional units

Domain	TESTCD	TEST	ORRESU	STRESU	CONU	CATEGORY	Data Source
LB	ALB	Albumin	gm/dL	gm/L	gm/dL	CHEMISTRY	LABS
LB	ALP	Alkaline Phosphatase	IU/L	microkat/L	IU/L	CHEMISTRY	LABS
LB	ALT	Alanine Aminotransferase	IU/L	microkat/L	IU/L	CHEMISTRY	LABS
LB	AST	Aspartate Aminotransferase	IU/L	microkat/L	IU/L	CHEMISTRY	LABS
LB	BASO	Basophils	10[3]/microL	10[9]/L	10[3]/microL	HEMATOLOGY	LABS
LB	BASOLE	Basophils/Leukocytes	%	%	%	HEMATOLOGY	LABS
LB	BILDIR	Direct Bilirubin	mg/dL	micromol/L	mg/dL	CHEMISTRY	LABS
LB	BILI	Bilirubin	mg/dL	micromol/L	mg/dL	CHEMISTRY	LABS
LB	BILIND	Indirect Bilirubin	mg/dL	micromol/L	mg/dL	CHEMISTRY	LABS
LB	BUN	Blood Urea Nitrogen	mg/dL	mmol/L	mg/dL	CHEMISTRY	LABS
LB	BUN	Blood Urea Nitrogen	mg/dL	mmol/L	mg/dL	CHEMISTRY	SLAB
LB	CA	Calcium	mg/dL	mmol/L	mg/dL	CHEMISTRY	LABS
LB	CL	Chloride	mEq/L	mmol/L	mmol/L	CHEMISTRY	LABS
LB	CO2	Carbon Dioxide	mEq/L	mmol/L	mEq/L	CHEMISTRY	LABS
LB	CREAT	Creatinine	mg/dL	micromol/L	mg/dL	CHEMISTRY	LABS
LB	CREAT	Creatinine	mg/dL	micromol/L	mg/dL	CHEMISTRY	SLAB
LB	EOS	Eosinophils	10[3]/microL	10[9]/L	10[3]/microL	HEMATOLOGY	LABS
LB	EOSLE	Eosinophils/Leukocytes	%	%	%	HEMATOLOGY	LABS
LB	GLUC	Glucose	mg/dL	mmol/L	mg/dL	CHEMISTRY	LABS
LB	HCT	Hematocrit	%	%	%	HEMATOLOGY	LABS
LB	HCT	Hematocrit				HEMATOLOGY	SLAB
LB	HCT	Hematocrit	%	%	%	HEMATOLOGY	SLAB
LB	HGB	Hemoglobin	gm/dL	gm/L	gm/dL	HEMATOLOGY	LABS
LB	HGB	Hemoglobin	%			HEMATOLOGY	SLAB
LB	HGB	Hemoglobin	gm/L	gm/L	gm/dL	HEMATOLOGY	SLAB
LB	HGB	Hemoglobin	gm/dL	gm/L	gm/dL	HEMATOLOGY	SLAB
LB	K	Potassium	mEq/L	mmol/L	mEq/L	CHEMISTRY	LABS
LB	LYM	Lymphocytes	10[3]/microL	10[9]/L	10[3]/microL	HEMATOLOGY	LABS
LB	LYMLE	Lymphocytes/Leukocytes	%	%	%	HEMATOLOGY	LABS
LB	MG	Magnesium	mEq/L	mmol/L	mEq/L	CHEMISTRY	LABS
LB	MONO	Monocytes	10[3]/microL	10[9]/L	10[3]/microL	HEMATOLOGY	LABS

SC | SE | SV | TA | TE | TI | TV | VS | ZE | TestCatg | QA_DATE

This information is used to verify the CDISC Controlled Terminology and the units of measurement provided to CRO.

Table – 4

Frequencies of key variable value are listed to verify counts

A	B	C	D	E	F
GRPID	**DSSCAT**	**DSCAT**	**DSTERM**	**DSDECOD**	**Frequency**
DS	COMPLETED	DISPOSITION EVENT	COMPLETED	COMPLETED	70
DS	COMPLETED STUDY MEDICATION	DISPOSITION EVENT	COMPLETED STUDY MEDICATION	COMPLETED	71
DS	DID NOT TAKE STUDY MEDICATION	DISPOSITION EVENT	DID NOT TAKE STUDY MEDICATION	OTHER	249
DS	DISCONTINUED	DISPOSITION EVENT	ADVERSE EVENT	ADVERSE EVENT	1
DS	DISCONTINUED	DISPOSITION EVENT	LACK OF EFFICACY	LACK OF EFFICACY	1
DS	DISCONTINUED	DISPOSITION EVENT	PROTOCOL VIOLATION	PROTOCOL VIOLATION	1
DS	DISCONTINUED	DISPOSITION EVENT	WITHDRAWAL BY SUBJECT	WITHDRAWAL BY SUBJECT	2
DS	DISCONTINUED STUDY MEDICATION	DISPOSITION EVENT	ADVERSE EVENT	ADVERSE EVENT	1
DS	DISCONTINUED STUDY MEDICATION	DISPOSITION EVENT	LACK OF EFFICACY	LACK OF EFFICACY	1
DS	DISCONTINUED STUDY MEDICATION	DISPOSITION EVENT	PROTOCOL VIOLATION	PROTOCOL VIOLATION	1
DS	DISCONTINUED STUDY MEDICATION	DISPOSITION EVENT	WITHDRAWAL BY SUBJECT	WITHDRAWAL BY SUBJECT	2
DS	EXCLUDED	DISPOSITION EVENT	ADVERSE EVENT	ADVERSE EVENT	1
DS	EXCLUDED	DISPOSITION EVENT	LOST TO FOLLOW-UP	LOST TO FOLLOW-UP	3
DS	EXCLUDED	DISPOSITION EVENT	WITHDRAWAL BY SUBJECT	WITHDRAWAL BY SUBJECT	14
Randomized Subj.					
75					
Unique					**Subcategory**

CO / DM / DS / EG / EX / IE / LB / MH / ORIGDM / PE / PR / QS / RELREC

REVIEW OF ADaM SETUP

The Clinical Data Interchange Standards Consortium (CDISC) Analysis Data Model (ADaM) is an emerging new industry standard for submitting analysis datasets to regulatory agencies, such as the U.S. Food and Drug Administration (FDA). While it provides several solutions for construction of analysis datasets, it is also subject to interpretation by users during implementation. The details of ADaM analysis variable naming conventions and usages are documented in the implementation guide and will not be repeated here.

ADaM IG presents metadata for two standard structures, as follows:

- ADSL – Subject Level analysis dataset, one record per subject
- ADXXX – Multiple-level-per-subject basic data structure

Metadata for the standard ADaM variables are presented in Section 3 of ADaM IG and the ADaM basic data structure and variables are discussed and illustrated in Section 4. The ADaM datasets displayed below are built from SDTM domains and naming convention followed the implementation guide.

1) Mapping verification

Below, three ADaM datasets are presented for illustration (See Tables 5a/5b – 7a/7b below). Due to space, only part of mapping specifications/ variables are shown associated with the datasets.

(a) ADSL is a subject level analysis dataset and it is the minimum requirement of ADaM. It contains important subject-level variables, such as population flags and treatment arms.

Table 5a
ADSL Specification

A	B	C	D	E	F	G
Variable	**Label**	**Type**	**Length**	**Controlled Terms or Formats**	**Source**	**Derivations**
STUDYID	Study Identifier	Char	18		DM.STUDYID	DM.STUDYID
ADDOMAIN	Analysis Domain Abbreviation	Char	7	"ADSL"	Derived	
USUBJID	Unique Subject Identifier	Char	18		DM.USUBJID	DM.USUBJID
SUBJID	Subject Identifier for the Study	Char	8		DM.SUBJID	DM.SUBJID
RANDDTC	Date/Time of Randomization	Char	19		DS.DSSTDTC DS.DSTERM	If DSTERM='RANDOMIZED' then RANDTC=DS.DSSTDTC
RANDDT	Date of Randomization, Num	Num	8	date9.		Derived from RANDDTC
RANDDTM	Date/Time of Randomization, Num	Num	8	datetime20.		Derived from RANDDTC
ARM	Description of Planned Arm	Char	20	Placebo Drug 2 mg Drug 5 mg Drug 10 mg	DM.ARM	DM.ARM is planned arm mapped to DM
				Not Treated		
TRTP	Planned Treatment Group	Char	20	Placebo Drug 2 mg Drug 5 mg Drug 10 mg	DM.ARM	DM.ARM
				Not Treated		
AGEU	Age Units	Char	6		DM.AGEU	DM.AGEU
AGEGRP	Age Group	Char	13	<65 >=65 and <75 >=75	DM.AGE	Derived from DM.AGE

ADAE / ADEX / ADCM / ADDEATH / ADDS / ADLB / ADMH \ ADSL / Macr

Table 5b
ADSL Analysis dataset

STUDYID	USUBJID	ARM	ACTLARM	RANDDTC	AGE	AGEU	AGEGRP	SEX	RACE	RACEGRP
9999-111	9999-111_9040-002	Drug 2 mg	Drug 2 mg	2007-01-11T09:29	70	YEARS	>=65 and <75	M	WHITE	WHITE
9999-111	9999-111_9040-003	Drug 5 mg	Drug 5 mg	2007-01-13T09:21	61	YEARS	<65	F	WHITE	WHITE
9999-111	9999-111_9040-004	Drug 10 mg	Drug 10 mg	2007-01-18T09:22	74	YEARS	>=65 and <75	M	WHITE	WHITE
9999-111	9999-111_9040-005	Placebo	Placebo	2007-02-05T15:30	83	YEARS	>=75	M	WHITE	WHITE
9999-111	9999-111_9040-006	Drug 2 mg	Drug 2 mg	2006-12-11T11:25	79	YEARS	>=75	M	WHITE	WHITE
9999-111	9999-111_9040-007	Placebo	Placebo	2007-01-05T10:02	51	YEARS	<65	M	BLACK OR AFRICAN AMERICAN	OTHER
9999-111	9999-111_9040-008	Drug 10 mg	Drug 10 mg	2007-02-02T10:25	63	YEARS	<65	M	WHITE	WHITE
9999-111	9999-111_9040-009	Placebo	Placebo	2006-10-20T10:45	66	YEARS	>=65 and <75	F	WHITE	WHITE

(b) ADLB - ADaM datasets should provide variables and metadata to fulfill below criteria:

- Identify observations that exist in the submitted study tabulation data (e.g. SDTM).
- Identify observations that are derived within the ADaM analysis dataset.
- Identify the method used to create derived observations.
- Identify observations used for analyses, in contrast to observations that are not used for analyses yet are included to support traceability or future analysis.

Some key variables from ADSL should be kept in the ADLB dataset but, since ADLB is usually a large dataset, there is a need to determine the optimal number of variables that are required and are analysis meaningful. A large dataset will degrade process performance. The variable length also plays an important part in the size of a dataset. The process performance section below will discuss this further.

Table 6a
ADLB Specification

A Variable	B Label	C Type	D Length	E Controlled Terms or Formats	F Source	G Derivations
STUDYID	Study Identifier	Char	18		ADSL.STUDYID	
ADDOMAIN	Analysis Domain	Char	7	"ADLB"	Derived	
USUBJID	Unique Subject Identifier	Char	18		LB.USUBJID	
PARAMCAT	Parameter Category	Char	30		LB.LBCAT	PARAMCAT=LB.LBCAT
ADT	Date of Specimen Collection, Num	Num	8	date9.	LB.LBDTC	Derived from LBDTC
ADTM	Date/Time of Specimen Collection, Num	Num		datetime20.	LB.COLDTM	Derived from LBDTC
PARAM	Parameter Description and Pref Units	Char	50		LB.LBTEST LBCONU	PARAM=LB.LBTEST || " (" || LBCONU || ")"
PARAMCD	Parameter Code	Char	8		LB.LBTESTCD	PARAMCD=LB.LBTESTCD
AVISIT	Visit Name	Char	20		LB.VISIT	AVISIT=LB.VISIT
AVISITN	Visit Number	Num	8		LB.VISITNUM	AVISITN=LB.VISITNUM
ANLFL	Analyzed Record Flag	Char	1	Y, null	LB.VISITNUM LB.LBTESTCD ADT	Sort by LB.USUBJID, PARAMCAT, LB.LBTESTCD, LB.VISITNUM, ADT. If FIRST.ADT or ABLFL='Y' then ANLFL='Y'
BASE	Result in Pref. Units at Baseline, Num	Num	8		AVAL	Sort LB data by USUBJID PARAMCAT, PARAM. If ABLFL='Y' then BASE=AVAL; output to temporary dataset, merge back with LB data by USUBJID, PARAM

ADAE / ADEX / ADCM / ADDEATH / ADDS / **ADLB** / ADMH / ADSL / Macr

Table 6b
ADLB Analysis dataset

STUDYID	USUBJID	PARAM	PARAMCD	PARAMCAT	AVISITN	ABLFL	AVAL	BASE	CHG	PCHG
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	1	Y	4.2	4.2	0	0
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	2		3.7	4.2	-0.5	-11.904762
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	3		3.3	4.2	-0.9	-21.428571
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	4		3.5	4.2	-0.7	-16.666667
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	5		3.3	4.2	-0.9	-21.428571
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	6		3.2	4.2	-1	-23.809524
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	7		3.2	4.2	-1	-23.809524
9999-111	9999-111_1001-001	ALBUMIN (gm/dL)	ALB	CHEMISTRY	14		3	4.2	-1.2	-28.571429
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	1	Y	20	20	0	0
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	2		14	20	-6	-30
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	3		15	20	-5	-25
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	4		17	20	-3	-15
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	5		53	20	33	165
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	6		44	20	24	120
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	7		33	20	13	65
9999-111	9999-111_1001-001	ALANINE AMINOTRANSFERASE (IU/L)	ALT	CHEMISTRY	14		25	20	5	25

(c) Time to event analysis of safety data - one record per subject per event of interest

Since most analysis datasets are derived from SDTM datasets, it is expected that there are some level of traceability between the SDTM dataset(s) and analysis dataset(s). In general, the CDISC ADaM standards recommend including as much supporting data in the traceability records and variables as possible, except in instances where it is not practical to do so such as eDiary data.

One key challenge in performing a time to event analyses is that some subjects may not experience the event of interest by the time that the study ends. For these subjects, the only information available is that the event did not occur within the duration of study. For these subjects, the time to event is said to be censored. There are other types of censored observations. For example, if a subject prematurely discontinues from study or experiences another type of event that prevents future assessment of the event of interest, the time to event for that subject would be censored at the time of discontinuation or occurrence of the specific event.

Table 7a
ADDEATH Specification

A	B	C	D	E	F	G
Variable	**Label**	**Type**	**Length**	**Controlled Terms or**	**Source**	**Derivations**
STUDYID	Study Identifier	Char	18		ADSL.STUDYID	
ADDOMAIN	Analysis Domain Abbreviation	Char	7	ADDEATH	Derived	set to 'ADDEATH'
USUBJID	Unique Subject Identifier	Char	18		ADSL.USUBJID	
CENSOR	Death or Censored Flag	Num	8	1 or 0	Derived	if ADSL.DTH180FL='N' then CENSOR=1 else CENSOR=0
PARAM	Parameter Description	Char	30	Time to Death (Days)	Derived	'Time to Death (Days)'
PARAMCD	Parameter Code	Char	8	DEATH	Derived	'DEATH'
AVAL	Analysis Value	Num	8		DTSTDT, ADSL.RANDDT	DTSTDT-RANDDT+1; if AVAL > 180 then change AVAL to 180
SRCDOM	Source Domain	Char	10	DS HO	derived	if CENSOR=0 then SRCDOM='DS' else if DTSTDTC=HO.HOENDTC then SRCDOM='DS'
SRCVAR	Source Variable	Char	10	DSSTDY HOENDY	derived	if CENSOR=0 then SRCVAR='DTDTHDTC' else if DTSTDTC=HO.HOENDTC then SRCVAR='DTSTDTC'
SRCSEQ	Source Sequence Number	Num	8		derived	if CENSOR=0 then SRCSEQ=DTSUPSEQ else if DTSTDTC=HO.HOENDTC then SRCSEQ=DTHOSEQ

ADAE | ADEX | ADCM | ADDEATH | ADDS | ADLB | ADMH | ADSL | Macr

Table 7b
ADDEATH Analysis dataset

STUDYID	USUBJID	CENSOR	EVENTDSC	PARAM	PARAMCD	AVAL	SRCDOM	SRCVAR	SRCSEQ
9999-111	9999-111_1001-007	1	CENSORED	Time to Death (Days)	DEATH	180	DS	DSSTDTC	3
9999-111	9999-111_1001-008	1	CENSORED	Time to Death (Days)	DEATH	180	DS	DSSTDTC	20
9999-111	9999-111_1001-009	0	DEATH	Time to Death (Days)	DEATH	86	DS	DTDTHDTC	500
9999-111	9999-111_1001-010	1	CENSORED	Time to Death (Days)	DEATH	180	DS	DSSTDTC	33
9999-111	9999-111_1001-011	1	CENSORED	Time to Death (Days)	DEATH	180	DS	DSSTDTC	465
9999-111	9999-111_1001-012	1	CENSORED	Time to Death (Days)	DEATH	180	DS	DSSTDTC	56
9999-111	9999-111_1001-013	1	CENSORED	Time to Death (Days)	DEATH	177	DS	DSSTDTC	78
9999-111	9999-111_1001-014	1	CENSORED	Time to Death (Days)	DEATH	180	DS	DSSTDTC	123
9999-111	9999-111_1001-015	1	CENSORED	Time to Death (Days)	DEATH	180	DS	DSSTDTC	334
9999-111	9999-111_1001-016	1	CENSORED	Time to Death (Days)	DEATH	61	DS	DSSTDTC	2

2) Analysis Data verification

For verification of ADaM data content, the approach is similar to the verification of SDTM discussed above. Since ADaM variables are standardized, different analysis datasets with common variable names can be displayed together. The reports (Table 8 -9) below summarize results from the integrated data source.

Table 8

Showing all datasets with PARAM/PARAMCD/RARAMN/PARAMCAT

Data Source	paramcd	param	paramn	paramcat
ADGLU	GLUC	Glucose (mg/dL)	100	GLUCOSE
ADGLU	HBA1C	Hemoglobin A1C (%)	250	GLUCOSE
ADGLU	INSULN	Insulin (microIU/mL)	150	GLUCOSE
ADLB	ALB	Serum Albumin (gm/dL)	2040	CHEMISTRY
ADLB	ALP	Serum Alkaline Phosphatase (IU/L)	2050	CHEMISTRY
ADLB	ALT	ALT (IU/L)	2030	CHEMISTRY
ADLB	AST	AST (IU/L)	2060	CHEMISTRY
ADLB	BASOLE	Absolute Basophil Count (cells/microL)	3040	HEMATOLOGY
ADLB	BILI	Total Serum Bilirubin (mg/dL)	2180	CHEMISTRY
ADLB	BUN	Blood Urea Nitrogen (mg/dL)	2010	CHEMISTRY
ADLB	CA	Serum Calcium (mg/dL)	2080	CHEMISTRY
ADLB	CL	Serum Chloride (mEQ/L)	2082	CHEMISTRY
ADLB	CRCLEST	Creatinine Clearance (mL/min)	2092	CHEMISTRY
ADLB	CREAT	Serum Creatinine (mg/dL)	2090	CHEMISTRY
ADLB	URATE	Serum Uric Acid (mg/dL)	2160	CHEMISTRY
ADLB	WBC	WBC Count (cells/microL)	3010	HEMATOLOGY
ADLIP	CHOL	Cholesterol (mg/dL)	800	LIPID
ADLIP	HDL	HDL Cholesterol (mg/dL)	820	LIPID
ADLIP	LDL	LDL-C (mg/dL)	830	LIPID
ADLIP	LDL-C	LDL-C (mg/dL)	830	LIPID
ADLIP	Non-HDL-C	Non-HDL-C (mg/dL)	840	LIPID
ADLIP	TG/HDL-C	Triglycerides (mg/dL) to HDL-C (mg/dL) R	850	LIPID
ADLIP	TRIG	Triglycerides (mg/dL)	810	LIPID
ADVS	BMI	BMI (kg/m[2])	1100	VITAL
ADVS	DIABP	Diastolic BP (mmHg)	1020	VITAL
ADVS	MABP	Mean Arterial Blood Pressure (mmHg)	1050	VITAL
ADVS	PULSE	Pulse Rate (beats/min)	1040	VITAL
ADVS	RESP	Respiration (breaths/min)	1060	VITAL
ADVS	SYSBP	Systolic BP (mmHg)	1030	VITAL
ADVS	TEMP	Temperature (C)	1070	VITAL

ADGLU / ADLB / ADLIP / ADSL / ADVS / TestCode / **TestCatg** / FixLength /

Table 9

Showing counts of different terms from studyid

STUDYID	SAFFL	Frequency	Cumulative Frequency
9999-111	N	14	14
9999-111	Y	3152	3166
9999-112	N	11	3177
9999-112	Y	548	3725
9999-113	N	6	3731
9999-113	Y	96	3827

Reported Term for the Adverse Event	Dictionary-Derived Term	Body System or Organ Class	Study Identifier	Stroke/Cerebrovascular Flag	Cnt
(L) KNEE ABRASION	EXCORIATION	INJURY, POISONING AND PROCEDURAL COMPLICATIONS	9999-111	N	1
(L) LOWER LIP ULCER	LIP ULCERATION	GASTROINTESTINAL DISORDERS	9999-113	N	1
(L) SHOULDER PAIN	MUSCULOSKELETAL PAIN	MUSCULOSKELETAL AND CONNECTIVE TISSUE DISORDERS	9999-111	N	1
(R) GROIN PAIN	GROIN PAIN	MUSCULOSKELETAL AND CONNECTIVE TISSUE DISORDERS	9999-111	N	1
A FIB	ATRIAL FIBRILLATION	CARDIAC DISORDERS	9999-111	N	1
ABDOMEN PAIN	ABDOMINAL PAIN	GASTROINTESTINAL DISORDERS	9999-111	N	1
ABDOMENAL PAIN	ABDOMINAL PAIN	GASTROINTESTINAL DISORDERS	9999-111	N	1
ABDOMINAL BLOATING	ABDOMINAL DISTENSION	GASTROINTESTINAL DISORDERS	9999-112	N	1

ADAE / ADCM / ADDEATH / ADDS / ADEX / ADLB / ADMH / ADSL / Test

3) Process performance

In order to ensure efficiency in running analysis programs, in addition to determining the right number of variables to keep, the length of each variable should be optimized since the integrated datasets are usually large. Cutting down on length will speed up running the jobs during development/debugging as well as supporting post-production agency requests. Please note that there are several ways to improve efficiency. However, since the CRO was responsible for the programming activities, this checking is one that the partner can easily identify and provide suggestions about.

Table 10 lists the ADaM datasets and their associated variable length. 'Max_Len' contains the actual maximum length of all values of a specific variable. 'Defined_len' has the assigned length of each variable. One can see that some variables length of 200 can really be shortened.

Table 10
Showing counts of different terms from studyid

ds_name	var_name	max_len	defined_len
ADAE	AEBODSYS	67	80
ADAE	AEDECOD	61	200
ADAE	AETERM	99	200
ADAE	CNTRYNAM	18	50
ADAE	STATUS	49	200
ADCM	CMDECOD	45	80
ADCM	CMDOSTXT	30	60
ADCM	CMDOSU	14	50
ADCM	CNTRYNAM	18	50
ADCM	STATUS	49	200
ADDEATH	CNTRYNAM	18	50
ADDEATH	DTDTHCSE	39	50
ADDEATH	STATUS	49	200
ADDS	CNTRYNAM	18	50
ADDS	DSDECOD	21	200
ADDS	DSSCAT	49	50
ADDS	DSTERM	49	200
ADDS	STATUS	49	200
ADLB	ASTALTFC	42	50
ADLB	CNTRYNAM	18	50
ADLB	LBCONRC	18	200

COMMUNICATION

In addition to the expectations, documentation, and requirements, it is important to communicate progress and concerns periodically to facilitate work and perform a timeline checkup. An issue log for each study was found to be helpful to document issues identified during the review process; this tool helps communicate the issues in detail to assist both sponsor and CRO in documenting their understanding. Recurring meetings help in addressing issues from different functional areas and assist in keeping the deliverables on track.

CONCLUSION

In summary, a successful filing package is a collaborative effort between the CRO and sponsor. The sponsor should provide clear expectations and all necessary information to help the CRO perform the SDTM mapping and set up of analysis datasets for analysis. If the data is expected to be loaded into the sponsor's internal database, certain guidelines and standards also need to be considered. It's expected that the CRO will follow their internal SOPs, understand the sponsor's expectation and raise any questions they may have to keep their activities on target. CDISC standardization helps achieve an overall structural consistency when data from multiple vendors with different data structures need to be pooled. Since CDISC standards are still evolving, these changes may impact internal policies and procedures developed to comply with the new standards by requiring adaptations. In a similar manner, the sponsor needs to establish clear roles and responsibilities with a CRO and be responsive to their needs for guidance and flexibility.

REFFERENCES

Version 3.1.1 (V3.1.1) of the CDISC Study Data Tabulation Model Implementation Guide
http://www.cdisc.org/models/sdtm/v1.1/index.html

ADaM Implementation Guide, Version 1.0 (ADaMIG v1.0) Draft
http://www.cdisc.org/models/adam/V2.1_Draft/index.html

SDTM Validation – how can we do it right?
http://www.cdisc.org/publications/euinterchange09/Session5_Track2/CDISC_SDTM_Validation_Submission.pdf

ACKNOWLEGEMENTS

The authors would like to thank the management team for their encouragement and review of this paper.

TRADEMARKS

CONTACT INFORMATION

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

Christine Teng Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc. Rahway, NJ 07065 Christine_Teng@Merck.com	sas Certified Advanced Programmer
Margaret Coughlin Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc. Rahway, NJ 07065 Margaret_Coughlin@merck.com	