Paper LS-244

SDTM What? ADaM Who? A Programmer's Introduction to CDISC

Venita DePuy, Bowden Analytics

ABSTRACT

Most programmers in the pharmaceutical industry have at least heard of CDISC, but may not be familiar with the overall data structure, naming conventions, and variable requirements for SDTM and ADaM datasets. This overview will provide a general introduction to CDISC from a programing standpoint, including the creation of the standard SDTM domains and supplemental datasets, and subsequent creation of ADaM datasets. Time permitting, we will also discuss when it might be preferable to do a "CDISC-like" dataset instead of a dataset that fully conforms to CDISC standards.

INTRODUCTION

CDISC is the Clinical Data Interchange Standards Consortium (www.cdisc.org), whose mission is to develop and support platform-independent data standards. While they have also developed other guidelines, such as for data collection (Clinical Data Acquisition Standards Harmonization, CDASH), this paper focusses on the Study Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) structures.

In simple terms, CDISC's goal is to provide a standard data layout, across different companies and different systems. In 2004, the FDA announced that they preferred to receive data in CDISC SDTM format. In September 2013, the FDA announced that they would require study data in conformance to CDISC standards "in the near future." As a result, there is more and more demand for CDISC-compliant datasets by the pharmaceutical and medical device companies. Furthermore, if individual studies use CDISC, it is relatively straightforward to combine the datasets across trials to do the Integrated Summaries of Safety and Efficacy (ISS and ISE) to support a New Drug Application (NDA).

At a recent (September 2014) seminar I attended, a CDISC representative indicated that the FDA was expected set a date (possibly by the end of 2014), and that any study beginning more than 2 years after that date would be required to submit datasets to the FDA in CDISC format. While CDISC format is not absolutely required for FDA submissions as of yet, it is certainly recommended. It can decrease FDA review time and subsequently decrease time to market. It is also cheaper to do CDISC initially, than to retroactively create CDISC datasets for past studies in preparation for an NDA filing.

Some of the advantages of CDISC are:

- Standardized dataset names and layout
- Standardized variable naming conventions
- Standardized calculations for things such as study day, percent change from baseline, and other common variables.

Once a programmer or statistician is familiar with the CDISC layout, it decreases the time to become familiar with a new study. If you start working on a project that you know is in CDISC, you automatically know that the lab dataset is called LB, and the numeric test result in standard units is LBSTRESN, and that your hemoglobin result will have LBCAT='HEMATOLOGY' and LBTESTCD='HGB' and LBTEST='HEMOGLOBIN', with standardized units in LBSTRESU.

This paper is not intended to be a comprehensive reference, but is instead meant to be an introductory overview to CDISC from a programming perspective, based on my personal experience working as both a statistician and a programmer for pharmaceutical companies and clinical research organizations. There are many more SDTM domains, and *many* more standardized variables, than are discussed herein. I have tried to focus on key variables in SDTM domains that are modified when creating ADaM datasets, and how those variables change between the two.

Key reference documents are given at the end of this document.

OVERVIEW

In simple terms, SDTM datasets are "raw" data and ADaM datasets are the analysis datasets. It is most efficient to create SDTM datasets, then ADaM datasets, then the display outputs (tables, listings, and figures [TLFs]) based on the ADaM datasets.

SDTM datasets, in general, have 1 record per subject per visit per assessment; for example, the dataset of laboratory results (LB) will have 1 record per subject per visit per laboratory test. The original laboratory results, from the lab, might be in the form of 1 record per subject per visit (with different variables for each test), which would need transposed to incorporate into the SDTM format, or may be in a SDTM format already (which is definitely easier to work with, in my

opinion). Other datasets might have one record per person (demographics, DM) or one record per subject, visit, questionnaire, and questionnaire item (for efficacy assessments).

Variable names are standardized, the possible values for those variables are often standardized using controlled terminology, and derivations may be standardized. For instance, in SDTM, original laboratory results, vital signs results, or questionnaire results will be in a variable called --ORRES, where "--" is the domain name: LB, VS, or QS respectively. Since values may be reported in different units (such as weight being in kilograms or pounds), standardized values are in --STRESC (for character values) and/or --STRESN (for numeric values). The subsequent ADaM datasets use Analysis Values (AVAL for numeric, AVALC for character), which - in the absence of any other derivations, imputations, etc. - are as simple as AVAL = --STRESN and AVALC = --STRESC. This makes it straightforward to produce tables, summarizing AVAL by timepoint and treatment group.

It is important to note that the same variable should have the same attributes across different datasets. VISIT should have the same length in all SDTM datasets, PARAM should have the same length in all ADaM datasets, and so forth.

SDTM

The SDTM model groups domains (datasets) into 3 general types:

- The *Findings* class captures the results of planned observations, such as ECG results or questionnaires.
- The *Interventions* class captures things that were done to subjects, such as treatments administered.
- The *Events* class captures things that occurred (such as adverse events or medical history) and protocol milestones (e.g., randomization and study completion).

In addition, the *Special Purpose Domain* category includes specific data structures that do not fit into the other categories, such as: demographics and subject visits. While there are additional special purpose domains, such as Trial Arms (TA) that provide descriptive information about the overall study, the scope of this paper is limited to datasets that contain information on individual subjects. While only a limited number of datasets are discussed herein, many things from laboratory results are applicable to vital signs, ECG, physical exams, questionnaires, etc.

The *Findings* class includes:

- Drug Accountability (DA), including dispense and return records
- Electrocardiogram (EG) captures ECG results
- Inclusion/Exclusion Criteria Not Met (IE) does not list all criteria for all subjects, just the criteria which were not met.
- Laboratory test results (LB) such as hematology, chemistry, and urinalysis (excluding pharmacokinetic or microbiology)
- Physical examination results (PE)
- Questionnaires (QS) provides a standard format for a variety of instruments, such as the SF-36.
- Vital Signs (VS), including blood pressure, temperature, height, and weight

The *Interventions* class includes:

- Concomitant Medications (CM) – this typically includes non-study medications and therapies, regardless of when they were taken (i.e., not just concomitant medications)
- Exposure Domains: Exposure (EX), which includes treatment administration but not dispense/return records

The *Events* class includes:

- Adverse events (AE)
- Clinical events (CE), which captures clinical events of interest that may not typically be classified as adverse events. For example, a cardiology study might look at clinical events such as heart transplant, hospitalization for heart failure, ventricular assistive device (VAD) insertion, myocardial infarction, stroke, cardiac death, and non-cardiac death.
- Disposition (DS), which captures not only study completion or early termination, but also protocol-defined time points such as informed consent, randomization, and entry into a long-term follow-up period.

GENERAL VARIABLE STRUCTURE FOR SDTM DATASETS

The SDTM Implementation Guide (SDTMIG) lists which variables are required, expected, or permissible (either optional, or whose requirement depends on study characteristics) for each domain. Variable names, labels, types (character or numeric) are specified. In some instances, controlled terminology, a codelist, or format is also specified.

A unique subject identifier, USUBJID, is required for all datasets containing subject data. It is important that this uniquely identifies a subject *across* studies. For instance, if data from multiple studies are combined for an ISS or ISE, there might be a subject 01-001 in more than one study, so USUBJID needs to have more information. While not strictly required to be so, it is typically of the form STUDYID-SITEID-SUBJID, where USUBJID = ABC123-01-002 indicates protocol # ABC123, site #1, and patient #2 within that site. This is the primary subject identifier used throughout the datasets. The study identifier (STUDYID, e.g. ABC123) and subject identifier for the study (SUBJID, e.g. 01-002) are also retained in most datasets. While it may be possible to use SUBJID instead of USUBJID when programming, I recommend using USUBJID for more robust programming.

The domain name (DOMAIN; the two-letter dataset name), STUDYID, and sequence number (--SEQ; a numeric identifier within each USUBJID and dataset) are required for each dataset as well. Different datasets have different required variables, as described in the SDTMIG. Typically, a list of controlled values is displayed in the SDTMIG by the name of an external codelist; clicking on that codelist in the current SDTMIG (version 3.2) will connect you to the online document (such as Controlled Terminology) containing all the SDTM codelists. While lists are generally comprehensive, I have occasionally had times when a list did not contain the specific test, etc. for that particular study. In those instances, I've created a value that matches the format of the existing values as closely as possible. It is recommended that controlled terminology be submitted in upper case text, other than exceptions identified in the SDTMIG (see Section 4.1.3.2). While in the past there has been a single controlled terminology, now a second document for controlled terminology for questionnaires and functional tests (such as the 6-minute walk test) is also available; links are available through http://cdisc.org/terminology.

Character date variables, designated by variables ------DTC, are always in ISO 8601 format. They are of the form 2014-07-28, which makes it easy to sort by that variable (unlike other character date formats). Date/time variables, or variables which may be date or date/time (such as adverse event start date/time, AESTDTC) are similarly in ISO 8601 format, of the form 2014-07-28T08:00. More details on using the ISO 8601 format in SAS can be found in Wilson's paper, "Harnessing the Power of SAS ISO 8601 Informats, Formats, and the CALL IS8601_CONVERT Routine" presented at PharmaSUG 2012. For simple cases, you can convert a complete date into the proper format with: `AESTDTC = put( startdate, yymmdd10.)` For partial dates, no dashes or "UNK" are displayed; a year-only date is "2014" and a date with only month and year is "2014-07."

Other common variable suffixes include:

-----CD: code (for example LBTEST is laboratory test and LBTESTCD is laboratory test code). This variable is typically no more than 8 characters long, and most often is controlled terminology. For example, hemoglobin always has LBTESTCD='HGB' and diastolic blood pressure always has VSTESTCD='DIABP'. Treatment arm codes (ARMCD and ACTARMCD, discussed in the following section), maybe 20 characters long.

-----FL: character flag variables include baseline record indicators (LBBLFL, VSBLFL) in SDTM or analysis flag variables (ANL01FL) in ADaM. Depending on the variable, possible values may be 'Y' or missing, or 'Y' or 'N'. ADaM also has numeric flag variables (-----FN), although they are not present in SDTM.

DEMOGRAPHICS DOMAIN (DM)

The first SDTM dataset I create for a new study is often the DM dataset, which strictly has one record per subject. In addition to the general variables required for all SDTM datasets, the demographics dataset will also have:

RFSTDTC, reference start date/time, is often the date/time of first exposure to study treatment, and required for all randomized subjects. In some cases, it might be the randomization date, or the informed consent date (in a single treatment arm study).

RFENDTC, reference end date/time, is typically the completion/discontinuation date, and often the date/time of last exposure. It is also missing for unrandomized subjects (in randomized trials), if they are present in the database (screen failures are often not entered in the database).

RFXSTDTC and RFXENDTC, date/times of first and last study treatments, were added in SDTMIG v3.2 and are equal to the earliest and latest date/times in the exposure (EX) dataset for that subject. RFSTDTC has typically been used to calculate Study Day variables (--DY) in ADaM datasets, but RFXSTDTC (if different than RFSTDTC) may be preferred. For instance, in a medical device study in which some patients received the device and other received standard care, RFSTDTC might be the informed consent date since only half of the patients would have a device implantation date. In general, I would recommend trying to match the Study Day calculations given in the protocol. That is not always possible,

as Study Day is calculated as (date-reference date) for dates before the reference date, and (date-reference date + 1) for dates after the reference date. Functionally, this means that there is no "Day 0" and days are numbered … -3, -2, -1, 1, 2, 3…. If the protocol numbering has the first day of treatment as "Day 0", it will be impossible to match.

ARM and ARMCD represent the planned treatment arm (i.e., treatment as randomized).
ACTARM and ACTARMCD represent the actual treatment arm (i.e., treatment as administered).

These will be the same unless the subject was accidentally given the wrong treatment assignment. In general, ARM and ARMCD should be used for all efficacy displays, and ACTARM and ACTARMCD for all safety displays (this is typically defined in the statistical analysis plan [SAP]). For crossover studies, ARM and ACTARM would have values such as "DRUGNAME 20 mg, PLACEBO 30 mg", with ARMCD and ACTARMCD values of "DG20-PB30". The treatments for specific periods (i.e. DRUGNAME 20 mg for the first treatment period) are split out into different variables in the ADaM datasets.

BASIC DATA STRUCTURE FOR THE *FINDINGS* DOMAINS

The standard SDTM (and ADaM) data structure is to have one record per subject per visit per assessment. For example:

LB:

USUBJID	LBDTC	LBCAT	LBTESTCD	LBTEST	LBORRES	LBORRESU	LBSTRESC	LBSTRESN	LBSTRESU
ABC123-01-002	2014-07-29	HEMATOLOGY	HCT	HEMATOCRIT	43	%	43	43	%
ABC123-01-002	2014-07-29	HEMATOLOGY	HGB	HEMOGLOBIN	13.2	g/dL	132	132	g/L
ABC123-01-002	2014-07-29	HEMATOLOGY	HGB	HEMOGLOBIN	13.2	g/dL	132	132	g/L

VS:

USUBJID	VSDTC	VSTESTCD	VSTEST	VSORRES	VSORRESU	VSSTRESC	VSSTRESN	VSSTRESU
ABC123-01-002	2014-07-29	HEIGHT	HEIGHT	190	cm	74.8	74.8031	IN
ABC123-01-002	2014-07-29	HR	HEART RATE	45	BEATS/MIN	45	45	BEATS/MIN
ABC123-01-002	2014-07-29	TEMP	TEMPERATURE	37.5	C	99.5	99.5	F

EG:

USUBJID	EGDTC	EGTESTCD	EGTEST	EGORRES	EGORRESU	EGSTRESC	EGSTRESN	EGSTRESU
ABC123-01-002	2014-07-29	QTC	QT INTERVAL, CORRECTED	400	msec	400	400	msec
ABC123-01-002	2014-07-29	QTCB	QTCB – BAZETT'S CORRECTION FORMULA	405	msec	405	405	msec

As you can see from those examples, the variable structure is similar across the different domains. Common variables are:

- --DTC displays the assessment date (and time, if recorded) in ISO 8601 format, as described earlier.
- --TEST and --TESTCD identify what laboratory test, ECG parameter, or other assessment is being performed, typically using controlled terminology. While there are instances in which controlled terminology is not yet present for an uncommon assessment, updated terminology is now available quarterly.
- --ORRES and --ORRESU contain original (character) results and associated units
- --STRESC, --STRESN, and --STRESU contain standardized results in character and numeric variables respectively, and standardized units. This is important because results need to be in the same units for summary tables, although they may be collected in different units (such as weight being collected in kilograms or pounds).
- --CAT is expected for LB, and has values of HEMATOLOGY, CHEMISTRY, URINALYSIS, etc. as shown in the controlled terminology. EGCAT and VSCAT (not shown) are permissible and could be used to group types of results together, although I have not used it to day.
- VISIT and VISITNUM (not shown above) display a standardized visit name, and an associated visit number.

This consistent naming scheme makes it easy to name variables consistently across domains and across studies, as well as simplifying the creation of subsequent analysis datasets.

SUPPQUAL DOMAINS AND CREATING ADDITIONAL DOMAINS

When variables that do not fit into the SDTM structure are needed, they can be included in either a SUPPQUAL dataset (such as SUPPDM, which has supplementary demographic information) or separate datasets. SUPPQUAL datasets are used when the variables are specifically associated with 1 or more records in the parent dataset. For instance, race is typically collected using several pre-specified options (White, etc.) plus "Other, specify", where "Other" can be checked

and then a description is provided in the "specify" field. In those cases, a SUPPDM dataset is necessary to hold the specific description. The SUPPMH dataset can be used to contain details about medical history that cannot be contained in the allowed variables in the MH dataset. SUPPEX holds additional information about exposure, and so on.

Each SUPPQUAL dataset contains 2 variables, IDVAR and IDVARVAL, that (in conjunction with USUBJID) indicate which record(s) in the parent dataset those supplemental records are associated with. In most cases, IDVAR = --SEQ (where -- is the parent dataset, such as MH) and IDVARVAL contains the value of that variable. So, if a subject had a record of past smoking history in MH with MHSEQ=6, the records in SUPPMH containing duration of the smoking use and how many packs a day were used would both have IDVAR="MHSEQ" and IDVARVAL="6". In some cases (very seldom, in my experience), other variables like --GRPID might be used instead of the --SEQ variable. For example, if a set of laboratory results has a comment from the central lab associated with it, stating that the sample was received slightly outside of the protocol-specified window but was analyzed anyway, a LBGRPID variable (group identifier) might be used to contain the identifier for that laboratory sample, and the SUPPLB dataset would contain a single record for that comment, with IDVAR='LBGRPID' and IDVARVAL='GWOJ19521'.

Domain names (i.e. SDTM dataset names) beginning with an X, Y, or Z are reserved for study-specific datasets. While most collected data belongs in a pre-specified domain, there are instances where data need to be in a dataset and doesn't fit anywhere. One relatively common (in my experience) time is when a sponsor wants an inclusion/exclusion listing that displays yes/no results for each and every criteria for each subject. Since the SDTM IE dataset is specifically Inclusion/Exclusion Criteria Not Met, it is not designed to have all of the "criteria met" results (i.e., inclusion criteria = yes, exclusion criteria = no). So, a new dataset is needed to contain that information. You might pick the first possible domain name (XA) or you might try to mimic "IE" and call it YE; anything between XA and ZZ is permissible. The dataset will use the standard pattern of SDTM variables. If the dataset is called XA, I would identify each record within USUBJID by XASEQ, group the inclusion vs. exclusion criteria using XACAT, identify individual questions by XATESTCD=INCL01 and XATEST=INCLUSION CRITERIA 01, and record results in XAORRES= Yes or No and XASTRESC=Y or N.

Another time that' I've had to create a new domain is for a cardiology study that tracked clinical events. Events such as heart failure, myocardial infarction, etc. are reported by the sites, and then adjudicated by an independent clinical event committee. Events are either verified as clinical events by the committee (and placed in the CE dataset) or determined by the committee to not have met that study's criteria for clinical events. So, we created a XE dataset, to mirror CE, to contain non-events (and, in practice, this contained events that had not yet been adjudicated as well, since the committee meets every few months). In general, new datasets should be created sparingly, since the vast majority of data can fit one of the existing frameworks.

After the SDTM datasets are created, ADaM datasets are used to create datasets with any calculations, imputations, or other derivations.

ADAM

Each study needs an ADSL dataset to include treatment assignments, reference dates, and other demographic and baseline characteristics that will be widely used in analyses and output displays. If merging a set of variables from the ADSL data onto the SDTM dataset is all that is needed to create the output displays for that domain (for example, medical history), my practice is to not create a separate analysis dataset (ADMH), but instead merge the ADSL variables onto the MH dataset within the table or listing program.

Other datasets require limited manipulation after merging the SDTM dataset with ADSL: ADAE typically requires the derivation of a treatment-emergent adverse event flag, and ADCM typically requires the derivation of a concomitant medication flag.

Many datasets (specifically those derived from SDTM datasets in the *Findings* class) might require further manipulation, such as the calculation of changes or shifts from baseline. Efficacy assessments might require imputation, calculations of subscale scores or total scores for individual assessments, and so forth.

As with SDTM, the Analysis Data Model Implementation Guide (ADaMIG) provides a great deal of information and is a necessary reference tool when developing ADaM datasets.

ADSL

ADSL is the subject-level dataset for ADaM. In simple terms, it should include virtually everything that you want to merge onto another dataset. Treatment groups, treatment start and stop dates, any indicator variables for subgroup analyses, and demographic variables should go here, *at the very least*. I have a current study in which the ADSL dataset has over 200 different variables, many of which are indicator variables. Required variables are: STUDYID, USUBJID, SUBJID, SITEID, AGE, AGEU, SEX, and RACE, all of which can be taken directly from the DM dataset. Throughout ADaM, it is important to remember that when ADaM datasets are created any variable imported directly from an SDTM dataset without changing its name (such as USUBJID) must also have the same attributes (label, length, type) and values as the variable in SDTM.

One of the key uses of the ADSL dataset is to provide the analysis population inclusion flags, to be merged into all the other datasets. Character flags (with values Y & N) are needed for each population defined in the SAP, and at least one must be present for each trial. Additional population flags may be added. Standard flags are: ENRLFL (enrolled population), RANDFL (randomized population), SAFFL (Safety population), ITTFL (ITT population), FASFL (Full Analysis Set population), PPROTFL (Per-Protocol population), and COMPLFL (Completers population). If the character flag is used, the corresponding numeric flag (----FN) can also be included.

The description of the planned treatment arm, ARM, is also required. While ACTARM (actual treatment arm) is not required, I would also include it; the current ADaMIG (v1.0) was published in 2009 and predates the inclusion of ACTARM in the SDTMIG, and I would anticipate that ACTARM will be included in later versions of the ADaMIG. I also include ARMCD and ACTARMCD for ease of programming.

TRTSDT, the date of first exposure to treatment, is required if there is an investigational product used in the study (unless TRTSDTM, a combined date/time variable, is used instead). Similarly, the date of last exposure, TRTEDT (or TRTEDTM) is also required in those cases. The date of the first exposure in each treatment period, TRxxSDT (or TRxxSDTM), and equivalent date of last exposure TRxxEDT (or TRxxEDTM) are required if more than one treatment period or more than one treatment period are present in the study, such as in a crossover study. In these cases, xx indicates 01, 02, etc. should be used as appropriate. It is important to note that these variables are *numeric*, whereas the source variables in SDTM (typically EX.EXSTDTC and EX.EXENDTC) are character.

Planned treatment (ARM) is used to create the TRTxxP variables (TRT01P, TRT02P, etc.). When a subject receives different treatments during different parts of the trial, TRT01P will capture what the treatment is for the first period, TRT02P for the second period, and so forth. In studies in which a subject just receives one treatment or one combination of treatments, TRT01P is used. TRTxxPN, the numeric representation of planned treatment assignment, has a 1:1 relationship to the values in TRTxxP. It is generally most convenient to assign TRTxxPN in the order that treatments will be shown in the final displays. For example, if listings are sorted by treatment then subject, it's helpful to assign TRT01PN as 1=Active, 2=Placebo so that it's straightforward to sort by TRT01PN then USUBJID. At least TRT01P is required. TRTxxA and TRTxxAN are the equivalent variables for actual treatment (and would be derived from ACTARM). TRTxxA is only required when there are instances in which the planned treatment and actual treatments differ for at least one subject. In practice, programming is typically done before data collection is completed, so there is the potential for a future treatment misallocation. I recommend including both TRTxxA and TRTxxAN in ADSL, since safety output are usually produced on actual treatment instead of planned treatment.

STUDY TREATMENT IN OTHER ADaM DATASETS

Those 4 sets of treatment variables (TRTxxP, TRTxxPN, TRTxxA, TRTxxAN) are used to derive the variables TRTP, TRTPN, TRTA, and TRTAN in most other ADaM datasets. The difference between these is, basically, that the latter set of variables indicates the treatment at the time of the record.

In the simplest case, assume we have a study where subjects are randomized to either active or placebo, receiving the first dose of study treatment at Visit 1 (on May 29th) and then take it daily until completing the study or withdrawing from it. For illustration, suppose vital signs are collected at Screening (pre-treatment) and Visit 2 only.

ADSL *(assuming numeric TRTSDT & TRTEDT are formatted as DATE9.)*

USUBJID	TRT01P	TRT01PN	TRT01A	TRT01AN	TRTSDT	TRTEDT
ABC123-01-002	DRUGNAME 20mg	1	DRUGNAME 20mg	1	29May2014	31Jul2014

ADVS

USUBJID	VISIT	TRTP	TRTPN	TRTA	TRTAN	PARAM	PARAMCD	AVAL
ABC123-01-002	SCREENING					SYSTOLIC BLOOD PRESSURE	SYSBP	120
ABC123-01-002	SCREENING					DIASTOLIC BLOOD PRESSURE	DIABP	90
ABC123-01-002	SCREENING					HEART RATE	HR	35
ABC123-01-002	VISIT 2	DRUGNAME 20mg	1	DRUGNAME 20mg	1	SYSTOLIC BLOOD PRESSURE	SYSBP	118
ABC123-01-002	VISIT 2	DRUGNAME 20mg	1	DRUGNAME 20mg	1	DIASTOLIC BLOOD PRESSURE	DIABP	85
ABC123-01-002	VISIT 2	DRUGNAME 20mg	1	DRUGNAME 20mg	1	HEART RATE	HR	38

As you can see, treatment variables are blank when the subject is not receiving that treatment, and TRTP=ADSL.TRT01P, etc. when the subject is in the treatment period.

For studies with sequential treatments, TRTP & TRTA represent the planned or actual treatment to be received during that specific period. For example, a subject in a crossover study might receive one randomized treatment for 3 weeks, have a 1 week washout period, then receive the other randomized treatment for 3 weeks. In such a case, ARM might be "DRUGNAME 20 mg, PLACEBO 30 mg" as in the earlier example. (A full description of the treatment assignments is found in dataset TA, Trial Arms; ARM & ACTARM may not be comprehensive descriptions.) In this case, ADSL would reflect:

TRT01P = "DRUGNAME 20 mg"

TRT02P = "PLACEBO 30 mg"

Then, each assessment in the ADVS would be assigned to a study period (APERIOD) based on the vital signs assessment date (ADT, taken from VS.VSDTC) compared to each treatment period's start and stop date (TR01SDT & TR01EDT, TR02SDT & TR02EDT). TRTP would then be set equal to TRT01P if the assessment was in APERIOD=1, and equal to TRT02P if the assessment was in APERIOD=2. If the assessment fell between the end of the 1st period and the beginning of the 2nd period (during the washout period), TRTP would be missing. TRTPN, TRTA, and TRTAN would be assigned similarly.

CONVERTING A *FINDINGS* DOMAIN INTO ADaM

There are a few general steps, in addition to determining study treatment and analysis periods, when creating a dataset like ADVS (vital signs) or ADQS (questionnaires). Not all of these are applicable to every dataset:

- Create initial ADaM dataset from SDTM dataset
- Create any imputed records (DTYPE=LOCF, etc.)
- Create any new results, such as calculating BMI (which could be derived in either SDTM or ADaM)
- Create analysis visits (AVISIT) if visit windowing is used
- Identify baseline value (mark with BASEFL=Y) and calculate change from baseline (CHG)
- Create any necessary flags

First, create the initial dataset. Some variables (USUBJID, STUDYID) are taken directly from the SDTM dataset. Others (TRTP and other treatment variables) are derived from the assessment date in the SDTM dataset and the subject-level variables (like TRTSDT) in ADSL. --TEST and --TESTCD are converted to PARAM and PARAMCD. --STRESN and --STRESC, the numeric and character results, are renamed to AVAL and AVALC (analysis values). The assessment date (--DTC), is converted to the numeric Analysis Date (ADT). Other variables may also be necessary, as described in the ADaMIG.

Then, extra records should be created as needed for imputation specified in the SAP. Imputed records are identified by the presence of a DTYPE (derivation type) variable with a non-missing value, such as "LOCF". Observed (un-imputed) data will have a missing value for DTYPE. Imputation is typically performed on efficacy datasets but not on safety datasets; I have included this step herein as these steps are generalizable to other Findings domains. These additional records may be created in different ways; one approach would be to use a similar method to the creation of the BMI records (below). If efficacy datasets require both imputation and the creation of subscale or total scores, it is important to create the imputed records prior to calculating the scores.

Next, any new results should be calculated. BMI (body mass index) is calculated as (weight in kilograms) / (height in meters)^2. In this particular case, I would start with the SDTM dataset (VS) and transpose the values (previously VSSTRESN, now AVAL) by subject and assessment time (date [VSDTC, now ADT] and/or timepoint [VSTPT, now ATPT]) to have weight and height on the same row. Since height is often collected at just the first assessment, that value would need carried forward to the other visits for the purpose of BMI calculations, but not retained as an observed height for those visits. I would calculate BMI from the standardized numeric values (converting units as necessary) and then transpose to return the data to the original layout of one record per subject per timepoint per assessment. The same method would be used to create records for total scores, subscale scores, and so forth in efficacy datasets.

Values are typically summarized by the nominal visit name (i.e., the name of the visit on the Case Report Form [CRF]). In some studies, each visit is assigned to a nominal visit name based on the Study Day of the visit in relation to the targeted study day for each visit. For instance, a study has scheduled visits every 2 weeks (Study Day 1, 15, 29, etc.) and visits should be performed within 2 days of the target date (so Day 15 ± 2, etc.). There might be out of window visits, there might be multiple visits within a window (subject had out of range lab values and came back to have lab samples pulled again; subject was sick at a visit and unable to receive study treatment so returned a day or two later; etc.). The later visit may be entered as an Unscheduled visit in the study database, since the scheduled visit name was already used. But the latter ("unscheduled") visit may have the results that should be counted as that scheduled visit for data summaries. (Listings will typically display the verbatim visit name, VISIT.)

This is addressed by the use of an Analysis Visit (AVISIT) variable. The algorithm for determining AVISIT is typically described in the analysis plan, but is often something like: If there is only 1 visit within the visit window, AVISIT=VISIT. If there is >1 visit within the window and one has results which are missing, not done, etc., assign the relevant AVISIT value to the one with non-missing results. If there is >1 visit within the window and both have non-missing results, use the one which is later chronologically. Not all records need an assigned AVISIT value; a visit which is not mapped to a scheduled visit name in that algorithm would have a missing AVISIT value.

Typically, the baseline value is defined as the last non-missing value prior to the first date/time that treatment was

administered (be sure to check the protocol or statistical analysis plan for the exact definition for a particular protocol). Baseline records (often, 1 record per PARAMCD per USUBJID) should be flagged with ABLFL=Y. Then, those records should be merged onto the other records to show baseline value. Often, baseline value is only defined for post-baseline values, so you may wish to define BASE (the baseline value) only for records on or after the first dose of study treatment (ADSL.TRTSDT). For those records, BASE is defined as AVAL where ABLFL=Y (within PARAMCD and USUBJID, or whatever other criteria as appropriate) and change from baseline (CHG) is defined as AVAL – BASE. If present, percent change (PCHG) is defined as CHG/BASE. It should be noted that baseline, and change from baseline, are not applicable to many datasets.

There are many standardized flag variables in ADaM datasets. The ANLxxFL variables (ANL01FL/ANL01FN, ANL02FL/ANL02FN, etc.) are probably the most common ones that I've used. One example might be a study that excluded assessments taken more than 5 days after the last dose of study medication from analyses, due to concerns that the pain-relieving capabilities of the drug would not last >5 days after the last dose, so all of the 'how do you feel?' assessments would not accurately reflect the treatment effect. All assessments within 5 days of last dose could have ANL01FL='Y' (after merging in TRTEDT from ADSL) and analysis tables could be produced only for records where ANL01FL='Y' throughout. A similar set of variables (CRITx, CRITxFL, CRITxFN) allows similar user-defined flags, where the actual criteria is given in CRITx. My understanding is that ANLxxFL variables are geared for use across multiple datasets, while the CRITx/FL/FN variables vary between datasets. For example, ADVS might have CRIT01='Systolic BP > 120 and Change from Baseline > 10' and CRIT01FL='Y', but ADLB might use CRIT01 for a completely different criteria.

ADAE AND ADCM

While the majority of ADaM datasets are based on SDTM datasets from the *Findings* class, two other common ADaM datasets (from the *Events* and I*nterventions* classes, respectively) are ADAE and ADCM. The analysis datasets for adverse events (ADAE) and non-study medications (ADCM) tend to require fewer derivations than those from the *Findings* class.

ADAE is typically necessary in order to determine which events are treatment-emergent adverse events (TEAE). TEAEs should be defined in the SAP, but are commonly defined as an adverse event beginning at/after the time of the first dose of study treatment. Studies often also have the caveat that the onset of the AE must be within a certain time period after the last dose of study treatment. If times of first and last doses and onset times of AEs are collected, it is possible to compare both dates and times when determining TEAEs. Otherwise, AEs starting on the same date as the first dose are typically considered TEAE. TEAEs are marked using TRTEMFL=Y. While it is possible to flag TEAEs in the SDTM datasets (using SUPPAE, where QNAM=AETRTEM), I have typically seen this created in ADaM datasets, since it is most often a derived variable.

Non-study medications (e.g., prior medications, concomitant medications, and post-study medications) are typically categorized as prior, concomitant, or follow-up in the ADCM dataset. It is important to note that, while the SDTM and ADaM domain names (CM and ADCM) indicate that they contain concomitant medications, other non-study medication is included as well. Start and end dates for study treatment (ADSL.TRTSDT and TRTEDT) are used in conjunction with non-study medication start and stop dates (CM.CMSTDTC and CMENDTC) to classify a medication as prior, concomitant, or follow-up. Medication start and end dates are often partial dates (i.e., subject has been taking heart medication since 2009). The general rule is, if a medication *might* be concomitant (based on partial start or end dates, or completely missing start or end dates), it is classified as concomitant; the specifics are generally specified in the SAP. There is not a standardized name for these flags; it might be helpful to use 3 flags (PREFL, CONFL, FUPFL, to match names of similar variables in ADAE) or one single variable with values of PRIOR, CONCOMITANT, or FOLLOW-UP. My personal preference is to use the 3 separate flags, to simplify programming of display outputs.

CONCLUSION

While I realize that this paper just barely touches on the tip of the iceberg, I hope that it has introduced you to some of the most common datasets, variables, and derivations used when creating SDTM and ADaM datasets. I strongly suggest printing out the SDTMIG and ADaMIG for review as you begin programming in CDISC. While these documents do not cover every eventuality, they cover the vast majority of them, and the patterns used for existing cases can be used to extrapolate a reasonable approach to the unusual cases that you may come across.

REFERENCES

The following links are current at the time of publication; all documents may be found through www.cdisc.org.

- Clinical Data Interchange Standards Consortium, Inc. (CDISC*). "Study Data Tabulation Model Implementation Guide: Human Clinical Trials v3.2".* 2013. Available at: http://cdisc.org/sdtm
- Clinical Data Interchange Standards Consortium, Inc. (CDISC*). "CDISC SDTM Controlled Terminology, 2014-09-26."* Available at: http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc

- Clinical Data Interchange Standards Consortium, Inc. (CDISC*). "Analysis Data Model (ADaM) Implementation Guide v1.0".* 2009. Available at: http://cdisc.org/adam
- Clinical Data Interchange Standards Consortium, Inc. (CDISC*). "Analysis Data Model (ADaM) Data Structure for Adverse Event Analysis v1.0".* 2012. Available at: http://cdisc.org/adam
- Wilson, Kim. "Harnessing the Power of SAS ISO 8601 Informats, Formats, and the CALL IS8601_CONVERT Routine". Presented at PharmaSUG 2012, Paper # DS22-SAS. Available at: http://www.pharmasug.org/proceedings/2012/DS/PharmaSUG-2012-DS22-SAS.pdf
-

CONTACT INFORMATION

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

Venita DePuy

Owner, Bowden Analytics

bowden.analytics@gmail.com