Sound SDTM, Sound ADaM -
Orchestrating SDTM and ADaM Harmonization

Soumya Rajesh, Nancy Brucken, and David Neubauer, IQVIA

ABSTRACT

Sound SDTM data is integral to having sound ADaM data. The ADaM model says "Whereas ADaM is optimized to support data derivation and analysis, CDISC's Study Data Tabulation Model (SDTM) is optimized to support data tabulation". Often times those who implement SDTM are not implementing ADaM and vice versa and they may not be working in harmony. If how the data will be analyzed has not been considered, tasks such as the definition of treatment arms and elements, and the assignment of collected data to SDTM domains can present various challenges during analysis and with traceability. This paper will cover some of the situations the authors have encountered and discuss how SDTM and ADaM implementation can be better in tune.

INTRODUCTION

A clinical study begins with questions to be answered, and forms to collect data to answer those questions. The forms are designed to be easily and correctly completed by sites, investigators and patients in order to increase the quality and accuracy of the collected data. SDTM standards then govern the mapping of that data into predictable storage locations. However, when weighing the sometimes-competing priorities of ease of collection and analysis readiness, the emphasis lands on the former due to the more severe consequences of collecting poor quality data. It is left to the biostatisticians and statistical programmers to reassemble the pieces for analysis. This paper provides some suggestions for structuring SDTM datasets to make this process easier.

TREATMENT ARMS, ELEMENTS

When creating SDTM Trial Design Model Datasets and their derived counterparts in the special purpose domains it's easy to focus on the rules and/or assumptions specific to those domains and variables. While adhering to those rules and assumptions is baseline conformance, one must balance the planned trial design, the actual execution of the study, and the study's analysis needs while working with this data.

A protocol study design schema and schedule of assessments can be interpreted in different ways by different people. One individual creating ELEMENT values in TA and TE domains might want to account for time between doses as rest elements, while another might see longer treatment elements with no defined/fixed rest element. On an oncology study some might interpret a dose as one treatment element, another might suggest a treatment element for each cycle, while a third may create one treatment element inclusive of all doses and cycles of said treatment. It really should depend on how the protocol was written and how the tabulated data will be analyzed. Recreating the exposure dataset in subject elements serves little purpose as it is redundant, and the data may not be analyzed at that granular a level. The SDTM should clearly and concisely describe the planned elements while allowing for the actual to be derived and also realizing some of this may have reuse in the ADaM ADSL treatment variables.

Be mindful to set values for ARMCD, ARM, ETCD, and ELEMENT so that you can try to convey some meaning of what is happening in that ARM or within that ELEMENT. An ARMCD = '1' may be easy to sort or query, but it conveys no meaning whereas using fragments of what appears in ARM can. ARMCD = 'A50OL' is a way to use fragments of the contents of ARM to still be short and easy to query for but also meaningful and descriptive in its value. It's also good to keep other studies that may be part of a pooled analysis in mind when setting ARM and ARMCD values. Reuse of ARMCD, ARM, ETCD, and ELEMENT values where applicable is good for comparing the same treatment assignments, but you should have enough specificity to distinguish unlike ARM values from each other. Doing so in SDTM will make it easier when ADaM programmers pool the ADaM data from multiple studies for an ISS or ISE.

Example: A relatively simple two arm open label comparator study where dosing is every day for 7 days. In TA, there is no need to have 7 repeating elements for treatment, instead 1 element per drug.
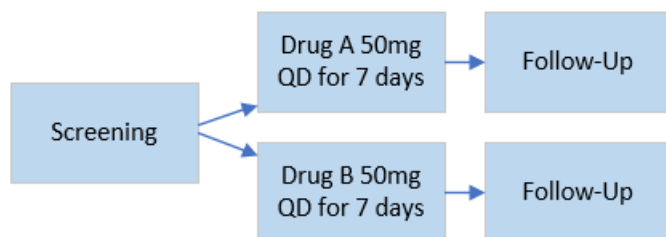
Figure 1: Study Schema

Row	STUDYID	DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	TATRANS	EPOCH
1	ABC-123	TA	A50OL	Drug A 50mg QD P7D Open Label	1	SCRN	Screening	Randomized to Drug A		SCREENING
2	ABC-123	TA	A50OL	Drug A 50mg QD P7D Open Label	2	A50QDOL	Drug A 50mg QD P7D Open Label			OPEN LABEL TREATMENT
3	ABC-123	TA	A50OL	Drug A 50mg QD P7D Open Label	3	FUP	Follow-Up			FOLLOW-UP
4	ABC-123	TA	B50OL	Drug B 50mg QD P7D Open Label	1	SCRN	Screening	Randomized to Drug B		SCREENING
5	ABC-123	TA	B50OL	Drug B 50mg QD P7D Open Label	2	B50QDOL	Drug B 50mg QD P7D Open Label			OPEN LABEL TREATMENT
6	ABC-123	TA	B50OL	Drug B 50mg QD P7D Open Label	3	FUP	Follow-Up			FOLLOW-UP

Table 1: ta.xpt

Provide clear and accurate descriptions of what is being performed in an element as well as the start rules. Elements having no gaps between them in the Subject Elements domain also means there should be no gaps in the elements in Trial Arms; one element's end becomes the next element's beginning so the conditions for that start and/or end rule need to account for many ways that element can actually end (not just the ideal planned condition).

Row	STUDYID	DOMAIN	ETCD	ELEMENT	TESTRL	TEENRL	TEDUR
1	ABC-123	TE	SCRN	Screening	Informed consent obtained	First dose of study drug	P14D
2	ABC-123	TE	A50QDOL	Drug A 50mg QD P7D Open Label	First dose of drug A	7 days after the start of element or treatment discontinuation	P7D
3	ABC-123	TE	B50QDOL	Drug B 50mg QD P7D Open Label	First dose of drug B	7 days after the start of element or treatment discontinuation	P7D
4	ABC-123	TE	FUP	Follow-Up	End of treatment	14 days after the start of element	P14D

Table 2: te.xpt

As with the trial design domain data being modeled in a way that is useful for analysis, ACTARM should be derived from actual treatment data but in a manner that is meaningful for analysis. While many think it is sufficient to inappropriately just set ACTARM = ARM, it can also be inappropriate to represent things that actually happened with too great a detail that the original intent is lost. There is little value in having overly rigid conditions in place for deriving ACTARM where many subjects are then represented as with a null ACTARM and ARMNRS = 'UNPLANNED TREATMENT' because they received slightly less than the target/planned dose when in ADaM the analysis would consider their ACTARM as not unplanned. It's best practice for those drafting SDTM specifications for ACTARM derivations to be aware of definitions in the statistical analysis plan and consult with biostatisticians and analysis programmers to ensure that

these types of derivations in SDTM retain their purpose in SDTM and planned reuse in ADaM. There should be balance and care taken when specifying the ranges of dosing when comparing subjects with ACTARM values describing different dose levels. The actual dosing frequency might need to be considered and/or the length of a grace period for any analysis of time-based components to ACTARM. In the TA example above, it might be the case for that study that ACTARM is derived based on whether the subject had a dose or not (exclusive to one drug or the other) and not 7 days of dosing.

All these things should be described in programming notes that can be represented in a reviewer's guide that shows how SDTM and ADaM accommodated planned versus actual and too strict or too loose versus what's necessary but still true for analysis.

REASSEMBLE RECORDS AFTER SPLITTING

Related data should be collected and tabulated in a manner that can reflect those relationships (across records or datasets). Many sponsors shy away from RELREC believing it too difficult to either represent, work with, or describe. Often data that truly should be represented in different domains in SDTM need to be merged back together in ADaM for analysis needs. Don't leave those data disconnected or model data in the wrong domains for fear of RELREC. Instead, use dataset or record level relationships and linking variables --LNKID and --LNKGRP (or other variables) where appropriate.

For example, the best response to a prior anti-cancer medication should not be stored in SUPPCM just because that is more convenient. Instead, response to treatment belongs in the RS domain and a RELREC from the collected prior anti-cancer medications can be established using CMLNKGRP and RSLNKID with the values being unique enough that they don't conflict with other records in the datasets (perhaps including a 'PRIORAC-'prefix to a record or line identifier to distinguish).

Row	STUDYID	DOMAIN	USUBJID	CMSEQ	CMGRPID	CMLNKGRP	CMTRT	CMCAT	CMSCAT
1	ABC-123	CM	001-001	1	FOLFOX-001	PRIORAC-FOLFOX-001	FOLFOX	PRIOR ANTI-CANCER	REGIMEN
2	ABC-123	CM	001-001	2	FOLFOX-001	PRIORAC-FOLFOX-001	LEUCOVORIN CALCIUM	PRIOR ANTI-CANCER	CONSTITUENT
3	ABC-123	CM	001-001	3	FOLFOX-001	PRIORAC-FOLFOX-001	FLUOROURACIL	PRIOR ANTI-CANCER	CONSTITUENT
4	ABC-123	CM	001-001	4	FOLFOX-001	PRIORAC-FOLFOX-001	OXALIPLATIN	PRIOR ANTI-CANCER	CONSTITUENT

Table 3: cm.xpt

Row	STUDYID	DOMAIN	USUBJID	RSSEQ	RSLNKID	RSTESTCD	RSTEST	RSCAT	RSSCAT	RSORRES
1	ABC-123	RS	001-001	1	PRIORAC-FOLFOX-001	BESTRESP	Best Response	RECIST 1.1	PRIOR ANTI-CANCER	PD

Table 4: rs.xpt

Row	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID
1	ABC-123	CM		CMLNKGRP	PRIORAC-FOLFOX-001	MANY	CM-RS-PRIORAC
2	ABC-123	RS		RSLNKID	PRIORAC-FOLFOX-001	ONE	CM-RS-PRIORAC

Table 5: relrec.xpt

Putting in the extra effort when designing forms, to capture identifiers or values that show the relationships that exist amongst data modeled in different SDTM domains, will ensure the relationships are not only collected and then tabulated but also easily merged and analyzed in ADaM.

DISEASE RESPONSE & CLINICAL CLASSIFICATIONS

The RS domain as we know it was previously designed to collect data pertaining to Oncology Disease Response to treatment, to connect with/represent the response evaluation(s) determined from the measurements or assessments in TR (Tumor Results) domain. Since the release of SDTMIG v3.3 however, the scope of this domain has expanded to also include the Clinical Classifications use case. It includes staging of the disease that's used in oncology and other therapeutic areas as well. The data collected in this domain may or may not be a part of a standard CRF – they could be based on an evaluator providing response evaluations based on published criteria. While the disease response use

case uses the ONCRSCAT codelist for RSCAT, the Clinical classifications use case uses CCCAT codelist for RSCAT. There are various supplements prepared for clinical classification instruments and therapeutic-area disease response criteria use cases in the QRS standards area of the CDISC website.

As we delve further into the domain, we'll discuss more on how to handle historical data, especially best response for prior chemotherapy regimens. Here are a few points to consider:

- Historical data is only split or distinguished for MH & AE. Everywhere else the data should be mapped where it goes – like CM (Prior Con Meds), Lab, VS, EG etc. Similarly historical disease response should be put into RS and SUPPRS, rather than in SUPPCM – for example "Best response for prior chemotherapy regimens".

- Many sponsors choose to put "Best response for prior chemotherapy regimens" in SUPPCM for ease of programming and to keep the data together with CM. However, that's not best practice or the ideal mapping strategy. Assumption #4 & 8 of the RS domain in the SDTM IG talk about how to use RS to map 'Best response'.

- Best response to the last chemotherapy regimen is often summarized in ADaM – in the form of a table for Disease response characteristics

- The main concern from the ADaM perspective is how to tie the data together from CM or PR to RS.

The RS domain disease response criteria use case may include records derived by the investigator or with a data collection tool, but not sponsor-derived records. Sponsor-derived records and results should be provided in an analysis dataset (ADaM). For disease response criteria, the BEST Response assessment records are included in the RS domain only when provided by the investigator or an independent assessor (i.e., Best responses that are derived by the sponsor for the analysis are not included in the RS domain). When using the RS domain to represent response evaluation or clinical classification instruments that incorporate data from other domains:

- In the oncology setting, the response to therapy would often be determined, at least in part, from data in the TR domain. Data from other sources (in other SDTM domains) might also be used in an assessment of response (e.g., lab test results, assessments of symptoms).

- Oncology response assessments sometimes include symptomatic deterioration.

Symptomatic deterioration may be considered as non-radiologic evidence of progressive disease. Symptomatic deterioration is recorded in RS with RSTEST = "Symptomatic Deterioration" and the standardized response (e.g., "PD") in RSSTRESC.

Best response, duration of response, or the progression to prior therapies and follow-up therapies may be represented in the RS domain by linking or relating the record in RS to a group of records or treatment regimens in Prior Concomitant Medications (CM) using CMLNKGRP and RSLNKID. Likewise, the link to Procedures (PR) (e.g., radiotherapy, surgery) would be made using PRLNKGRP. Here are a couple of examples to show 'best response' and the 'overall response' of progression to prior and follow-up therapies:

Row	STUDYID	DOMAIN	USUBJID	RSSEQ	RSLNKID	RSTESTCD	RSTEST	RSCAT	RSSCAT	RSORRES	RSSTRESC	VISITNUM	RSDTC
1	ABC-123	RS	001-001	1	PRIORAC-FOLFOX-001	BESTRESP	Best Response	RECIST 1.1	PRIOR ANTI-CANCER	PD	PD	1	
2	ABC-123	RS	001-001	1		OVRLRESP	Overall Response	RECIST 1.1	ON-STUDY	CR	CR	12	2010-04-02

Table 6: rs.xpt

How does the Clinical Classifications use case differ from the Disease Response use case in RS? While Best response and Overall response would map under the Disease Response use case in RS, TNM Classification or any such classification of disease staging would map under the Clinical Classification use case of the RS domain. This is because Clinical Classifications are measures that serve as a surrogate for, or ranking of, disease status (such as disease staging), symptoms, or other physiological or

biological status. The AJCC TNM staging system is a common cancer staging system. It looks at T (tumor expansion), N (extent of cancer spread to lymph nodes) and M (metastases or spread of cancer to other organs). It also includes the transition from non-metastatic to metastatic colorectal cancer, where the cancer has spread through the colon wall and may have spread to nearby organs or lymph nodes.

Here is an example of how TNM staging would be represented in SDTM based on the initial diagnosis of breast cancer: Again, the RELREC serves as a guide for reassembling the components.

Row	STUDYID	DOMAIN	USUBJID	MHSEQ	MHLNKID	MHTERM	MHDECOD	MHEVDTYP	MHCAT
1	ABC-123	MH	001-001	1	BR-CA-01	BREAST TUMOR MALIGNANT	Breast cancer	INITIAL DIAGNOSIS	PRIOR CANCER HISTORY

Table 7: mh.xpt

Row	STUDYID	DOMAIN	USUBJID	RSSEQ	RSLNKGRP	RSTESTCD	RSTEST	RSCAT	RSSCAT	RSORRES	RSSTRESC	VISITNUM	RSDTC
1	ABC-123	RS	001-001	1	BR-CA-01	AJCC101	AJCC1-Primary Tumor (T)	AJCC V7	PRIOR CANCER HISTORY	T1	T1	1	2010-10-26
2	ABC-123	RS	001-001	2	BR-CA-01	AJCC102	AJCC1-Regional Lymph Nodes (N)	AJCC V7	PRIOR CANCER HISTORY	N1	N1	1	2010-10-26
3	ABC-123	RS	001-001	3	BR-CA-01	AJCC103	AJCC1-Distant Metastasis (M)	AJCC V7	PRIOR CANCER HISTORY	M0	M0	1	2010-10-26

Table 8: rs.xpt

Row	STUDYID	RDOMAIN	USUBJID	IDVAR	IDVARVAL	RELTYPE	RELID
1	ABC-123	MH		MHLNKID	BR-CA-01	ONE	MH-RS-PRIORAC
2	ABC-123	RS		RSLNKGRP	BR-CA-01	MANY	MH-RS-PRIORAC

Table 9: relrec.xpt

SINGLE OR SPLIT FA DATASETS?

The purpose of Findings About domains is to store findings related to Events or Interventions that may or may not be related to other datapoints on the CRF – as in stand-alone relational data. CDISC standards and the FDA allow sponsors to choose from different ways to represent Findings About data collected in the study. However, we want to make sure that either approach is solid in maintaining traceability – whether it be a single FA dataset or split datasets. A single FA dataset may not be the best approach if the data pertains to different contexts or parent domains in a study. In such cases, splitting the FA domain into physically separate datasets following the guidance described in Section 4.1.6, 'Additional Guidance on Dataset Naming' might work best, wherein you would have to assign a unique 2-character domain code suffix followed by the FA domain. For example, if a study collects Findings About Clinical Events, then a dataset called FACE would be created, or Findings About Medical History would reside in FAMH. When choosing either approach we want to keep in mind the following points:

- If representing FA as a single domain, then separate the datapoints by different FACAT and/or FASCAT values.

- If representing a split FA domain, then make sure to follow the guidance in 'Section 4.1.7: Splitting Domains'.

Splitting the FA domain into topic-based datasets like FACE and FAMH would be especially beneficial for downstream processes like ADaM datasets because the programming would be much easier, and traceability would be cleaner. On the other hand, if there are complex inter-domain relationships between FA and multiple parent domains then it may be better to leave the data in a single dataset separated by --CAT or --SCAT and --OBJ. ADaM would probably summarize by reaction, by timepoint, by severity; and go into FACE for the details. However, this would really depend on the therapeutic area or indication of the study. There is no 'one rule fits all' solution. If your analysis or ADaM datasets requires the information in one dataset versus split datasets, then you would have to group them together. The main concern however is to put data in the right place. If your organization prefers to put all the FA data in one dataset then make sure you have the right qualifier variables: --CAT, --SCAT, --OBJ and maybe identifiers like --GRPID / --SPID / --LNKID – so that the data can be easily split out into multiple analysis datasets as needed.

–ORRES/--STRESN/--STRESC

For data to be summarized across subjects and visits, and to enable calculation of changes from a baseline value, each test or parameter must be expressed in a consistent set of units throughout a study. Much of the world has adopted SI (International System) units when reporting clinical laboratory data, while U.S. labs may report values in a mixture of SI and U.S. "conventional" units. The FDA has recommended that sponsors contact their reviewers prior to submission in order to reach agreement on displayed units for important lab tests.

In order to provide the best communication between SDTM and ADaM, --STRESN values should reflect the units that have been agreed upon by both sponsors and regulatory agencies. --ORRES values can then contain the values in the units in which they were originally collected. That convention provides the greatest amount of traceability between what was collected and what was reported, and allows the values of --STRESN to be copied directly to AVAL in an ADaM Basic Data Structure (BDS) when that test is used for analysis.

Here's an example of such an approach. In this study, glucose is supposed to be summarized in units of mmol/L. However, as you can see, it has been collected in several different sets of units.

Row	LBTESTCD	LBORRES	LBORRESU	LBSTRESN	LBSTRESU
1	GLUC	5.2	mmol/L	5.2	mmol/L
2	GLUC	80	mg/dL	4.4	mmol/L
3	GLUC	740	mg/L	4.1	mmol/L

Table 10: Sample Lab Values

This structure also allows a reviewer to explore the SDTM dataset directly, because the LBSTRESN values are all on the same scale across subjects and visits.

Questionnaire, ratings and scale (QRS) data may be handled slightly differently. With QRS items, --ORRES represents the text that is displayed on the entry form. --STRESN represents the coded value used for scoring, and –STRESC represents the character equivalent of that coded value. If any of the coded values need to be reversed or otherwise modified for scoring purposes, a decision must be made as to whether that will happen when mapping the SDTM datasets, or during the programming of the ADaM datasets, and then be documented accordingly. In general, --STRESN should be copied into AVAL and –ORRES into AVALC when creating the ADQRS analysis dataset, so long as there is a 1:1 relationship between the two values within a parameter. Populating both AVAL and AVALC on the same record is acceptable in this situation because there is an inherent code-decode relationship between the two values. If there are any changes in value between –STRESN and AVAL, --STRESN should be carried over into the analysis dataset for traceability.

AE AND CM CODING

There can be considerable variation in how adverse events (AEs) and prior and concomitant medications are recorded by the sites, as there are often many ways to report the same event or medication. For example, a headache may be described by a patient as "head pain", "occasional headache", "pain at the top of head", or other similar phrases. However, because of this variation, a table summarizing AEs or medications by their originally recorded terms would be very difficult to interpret. Instead, we summarize on the coded terms by assigning the verbatim AE and medication text to coded values using a common dictionary, generally MedDRA for AEs and the WHODrug dictionary for medications, and then summarizing the coded terms instead.

To ensure the traceability of AE and medication records from the tables back to SDTM, the SDTMIG indicates that the AE dictionary preferred terms and body system organ classes (SOC) used for analysis be mapped to the AEDECOD and AEBODSYS variables, respectively. AESOC is used to store the primary SOC and will be identical to AEBODSYS if the primary SOC is summarized. The full MedDRA hierarchy of terms has been classified as expected, so a complete set should be included in the SDTM AE dataset.

Similarly, concomitant medications are summarized by preferred term and often by medication class, with the coded terms again used to minimize variation in how medications are described. CMDECOD should contain the preferred medication term displayed on the tables. If medications are also grouped by classification or ATC code, CMCLAS and CMCLASCD should also be included in the CM dataset and contain the summarized values. ATC codes can be reported at different levels for different medications, so it's not unusual to find some medications where CMCLAS contains a level 4 code, and others where CMCLAS contains a level 2 or 3 code. Again, having the coded terms displayed on the tables stored in SDTM helps provide traceability between the tables and the collected data. In addition, statisticians and programmers are not equipped to code AEs and medications if the coded terms to be summarized are not included in SDTM.

CONCLUSION

As stated at the beginning of this paper, having sound SDTM is key to having sound ADaM. Employing techniques such as the use of linking and grouping variables to allow for easy merging of datasets, creation of RELRECs to explicitly describe relationships between datasets, and designing trial design datasets and assigning ARM and ACTARM values with input from biostatisticians and statistical programmers will go a long way towards the creation of harmonized SDTM and ADaM datasets in your clinical studies.

REFERENCES

SDTM IG v3.4. Available at
https://www.cdisc.org/system/files/members/standard/foundational/SDTMIG%20v3.4-FINAL_2022-07-21.pdf

ADaM IG v1.3. Available at
https://www.cdisc.org/system/files/members/standard/foundational/ADaMIG_v1.3_Release_Package_2.zip

CONTACT INFORMATION

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

Soumya Rajesh
IQVIA
soumya.rajesh@iqvia.com

Nancy Brucken
IQVIA
nancy.brucken@iqvia.com

David Neubauer
IQVIA
david.neubauer@iqvia.com