PharmaSUG China 2019 - Paper DS-068

How to ensure quality in data submission

Angelo Tinazzi, Cytel Inc.

ABSTRACT

Efficacy and safety of your drug are what matter, but lack of traceability, or poor or insufficient documentation might trigger questions and concerns. While this might not impact the final outcome of your submission, approval could be delayed if the reviewer starts questioning what you have done by requesting changes, or new deliverables to clarify aspects that were not sufficiently clear in your submission.

In my current positions as CDISC SME I'm exposed to several CDISC submission packages, either produced by my company or by other CROs or sponsors, for both SDTM and ADaM. In this capacity I have seen several define.xml and reviewer guides that demonstrate how differently individual users and companies approach the same mapping issue (SDTM) or same analysis "modelling" (ADaM). I have also observed wide variations in the level of details provided for example in a reviewer's guide, or a computational algorithm used to describe a derivation.

With this presentation I would like to share some of the main CDISC "Nonsense" I have seen in the CDISC packages I have reviewed; this can range from "nonsense" questions to complete misunderstanding of the CDISC IGs or bad documentation. I will share some of the feedback we have received or our sponsors received when making mock submission to the FDA demonstrating how the agency cares about details and ultimately in precision and cure for details in your submitted datasets and documentation.

INTRODUCTION

Over the last 5-10 years I have been exposed to a number of data submission packages, either created by my colleagues at Cytel or by the sponsor itself or by their CROs. Although this is continuously getting better, I can see in these packages there is still a lack of clarity or poor attention to details.

WHAT DO WE MEAN BY "QUALITY" IN DATA SUBMISSION?

Any "piece" submitted to the Health Authorities (HA) should be of good quality. Quality must be guaranteed not only in the submitted data and in the analysis results, but also in the documentation we provide in support of submitted data and analysis results. Attention should be payed to completeness and clarity of such a documentation as the reviewer at the HA needs to be able to understand what you have done.

EXPERIENCE GATHERED WHILE SUBMITTING DATASETS TO THE FDA

With the FDA you have the possibility of making a test submission, or *mock submission*, of your data submission package[1]. This usually includes only one study and possibly real data (making a mock submission with test data only does not make sense at all). Which study you use to make the test submission does not matter to the FDA as this is a pure technical test. However, we do recommend using a study with some sort of complexities in the way data are transformed into SDTM for example, if the study had several instances where the standard was "deviated". This is really your opportunity to seek technical advices from the technical team at the FDA and very often in our experience we did receive some good technical advice on technical aspects of for example define.xml we were not fully aware.

From the mock submission we did with our sponsor at Cytel or the sponsor did on its own, we have observed the level of details the FDA is seeking in your data submission package. For example the list of

[1] https://www.fda.gov/drugs/electronic-regulatory-submission-and-review/submit-ectd-or-standardized-data-sample-fda

key variables indicated in your define.xml, or the way you have documented in the reviewer guide (or justified) an unresolved conformance issue, or the computational algorithm you have described in your define.xml, variables in the define.xml that should have a controlled terminology assigned and they have not, improper use of variables in ADaM, incomplete reviewer's guide documentation such as not providing a rationale for all warnings and errors P21 reported during the validation of your package

All these issues are the object of FDA concerns and I recommend you correct them.

FDA HAS ITS OWN PREFERENCES

Regardless if you are working or not on a data submission project, if you apply CDISC standards you should be also aware of the additional HA requirements, either FDA or PMDA, with regards to data standards submission. These two HA have developed a strong process to help the sponsors submitting data in the most standardized way. As such, the two HA have developed a set of ad-hoc guidance where they clarify some aspects of the standards where different interpretations can be made or more than one option can be chosen. Awareness of HA data submission requirements should be "a must" in the curriculum of modern Statistical Programmers (but similarly for Biostatisticians)!

This unfortunately, is often not the case, notably for the following aspects:

- electronic Common Technical Document (eCTD) structure: where do our datasets go?
- naming conventions and limitations of filename length
- do I need to submit my SAS programs, and if yes, do they need to be executable?
- different requirements between FDA and PMDA with regards to naming conventions for reviewer guides
- recommendations for ADaM datasets creation

As mentioned above, the HA might also have specific requirements detailed in their guidance, for example when implementing SDTM. In some cases, the guidance simply provides the HA recommended approach when SDTM gives more than one implementation option; take for example the handling of the term OTHER where the FDA does recommend "*to map a collected value to OTHER when there is a controlled term available to match the collected value – even when the terminology allows for Sponsor expansion*", while the SDTM Ig gives at least three options from which the sponsor could choose.

In some circumstances, the HA guidance "contradicts" the standard. This is the case in SDTM for subjects randomized to the treatment but not treated (in DM). While the SDTM Ig (v3.2) recommends assigning ACTARMCD= NOTTRT, the FDA Study Data Technical Conformance Guidance requires ACTARMCD to be left blank. Of note, with the latest SDTM Ig (v3.3) this discrepancy has been removed.

Last, but not least, people often ignore that the data standards requirements, as it is meant by the HA such as the FDA, it is not just about CDISC: the eCTD previously mentioned, the medical dictionaries and controlled terminology are other standards to consider, as well as the requirements for PDF files or any other standard mentioned for example in the FDA standards catalogue.

There are also situations where statements in the HA guidance are the object of misunderstanding. For example, while I was finalizing my paper a sponsor came to me with a comment related to the CDISC package we did deliver for their submission:

> "*Our CDISC expert says that if we submit SDTM FDA wants us to also submit an ADSL regardless if ADaM was created or not in support of the Individual Clinical Study Report*"

The FDA Study Data Technical Conformance Guidance (SDTCG) has the following sentence:

> "*All submissions <u>containing standard analysis data</u> should contain an ADSL file for each study*"

The sentence is clearly saying the requirement applies to only studies for which we submitted an ADaM package; for example, in many submissions, for some legacy studies, we do submit only the migrated SDTM because these studies contributed to the Integrated Summary of Safety (ISS). The source and responsible of this possible misunderstanding is the FDA itself, because the way this requirement was “formulated” in the FDA Technical Rejection Criteria for Study Data could be the cause of such misunderstanding:

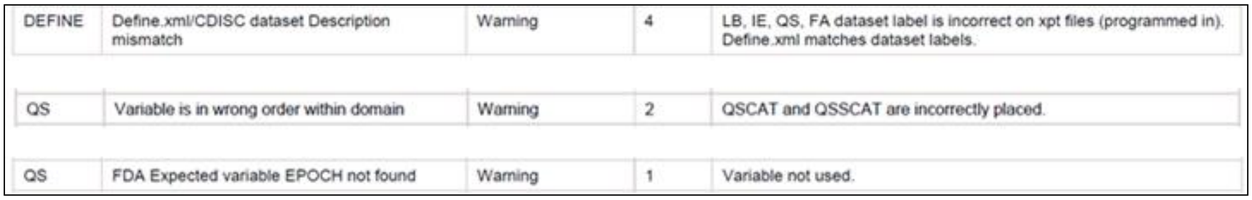
> “*DM dataset and define.xml must be submitted in module 4, sections 4.2.3.1, 4.2.3.2, 4.2.3.4. DM dataset, ADSL dataset, define.xml must be submitted in module 5, sections …..*”

We hope FDA will clarify or rephrase the sentence in future versions of the Technical Rejection Criteria document. Of course, it would not make sense to submit an ADSL if the data were not analysed using ADaM standards and if the study data were submitted in SDTM format for the only purpose of being integrated for example in a pooled ADaM for supporting an ISS.

DETAILS MATTER

In my professional career I have been exposed to several studies requiring the use of CDISC standards, either as a programmer study lead or as a CDISC SME reviewing CDISC packages. In this capacity, I have seen several define.xml and reviewer guides that demonstrate how differently individual users and companies approach the same mapping issue in SDTM or the same analysis "modelling" in ADaM. I have also observed wide variations in the level of details provided for example in a reviewer’s guide, or a computational algorithm used to describe a derivation in an ADaM define.xml.

Very often people pay different "attention" to details, perhaps (incorrectly) thinking they do not matter. If for example you created an SDTM dataset and for one variable you assigned the wrong label (not as per the SDTM Ig), or you did not assign a label at all and this has been detected from the Pinnacle21 validation, why not correct the issue in the dataset instead of leaving the issue as it is and justifying it in the study data reviewer guide as a "Programmatic Error"?



DEFINE	Define.xml/CDISC dataset Description mismatch	Warning	4	LB, IE, QS, FA dataset label is incorrect on xpt files (programmed in). Define.xml matches dataset labels.
QS	Variable is in wrong order within domain	Warning	2	QSCAT and QSSCAT are incorrectly placed.
QS	FDA Expected variable EPOCH not found	Warning	1	Variable not used.

Figure 1: Clinical Study Data Reviewer Conformance Section with by “bad” justifications of errors and warnings

As previously discussed we see more and more, from the mock /test submissions we do for the FDA or that our clients do with their CROs, that agency concerns about such details is increasing. You may think these are minor issues because they do not ultimately impact any result. However, you are risking your credibility with the FDA reviewer, who may conclude that your package is not of good quality. Furthermore, if you don’t address these issues on-time, you might receive a request from the FDA to correct them when you think you are done and your package is ready to be delivered. Fixing, for example, an SDTM dataset could have a "cascade" effect as you might also need to re-run other datasets or even re-generate the ADaMs (and outputs as well).

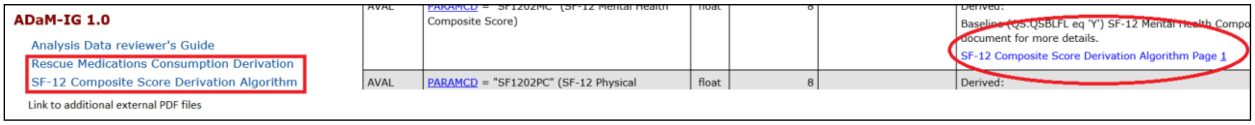
CREATING PROPER DOCUMENTATION

Alternatives to ADaM define.xml when describing complex algorithms

Proper documentation is of vital importance and could be a success factor for your submission. Do not cut corners! Try to imagine that you are the "recipient" of a data submission package and check, for example, if the explanation of a derivation in the define.xml is clear enough. Do not hesitate to find alternative ways if you identify that define.xml is not the most appropriate “tool” to describe a complex derivation. Question yourself for example if you need more than 1000 characters to describe your derivation, whether you might instead describe your complex algorithm in a separate document (e.g. a PDF document that you hyperlink in the define.xml)[2].

> *Do not cut corners! Try to imagine that you are the "recipient" of such package and check, for example, if the explanation of a derivation in the define.xml is clear enough*

See for example figure 2. This is from a define.xml of a successful FDA ISS/ISE submission, where the study had several endpoints related to different quality of life (QoL) aspects. QoL analysis often requires combining answers of several QoL questions into one summary score, by using complex algorithms requiring several steps and transformations e.g. normalizing the answer.



ADaM-IG 1.0

Analysis Data reviewer's Guide
Rescue Medications Consumption Derivation
SF-12 Composite Score Derivation Algorithm

Link to additional external PDF files

AVAL	PARAMCD = "SF1202MC" (SF-12 Mental Health Composite Score)	float	8		Derived: Baseline (QS.QSBLFL eq 'Y') SF-12 Mental Health Compo document for more details. SF-12 Composite Score Derivation Algorithm Page 1
AVAL	PARAMCD = "SF1202PC" (SF-12 Physical	float	8		Derived:

Figure 2: An ADaM define.xml referencing an external PDF file containing desctiption of a complex algorithm

SDTM define.xml should not contain legacy datasets migration specifications

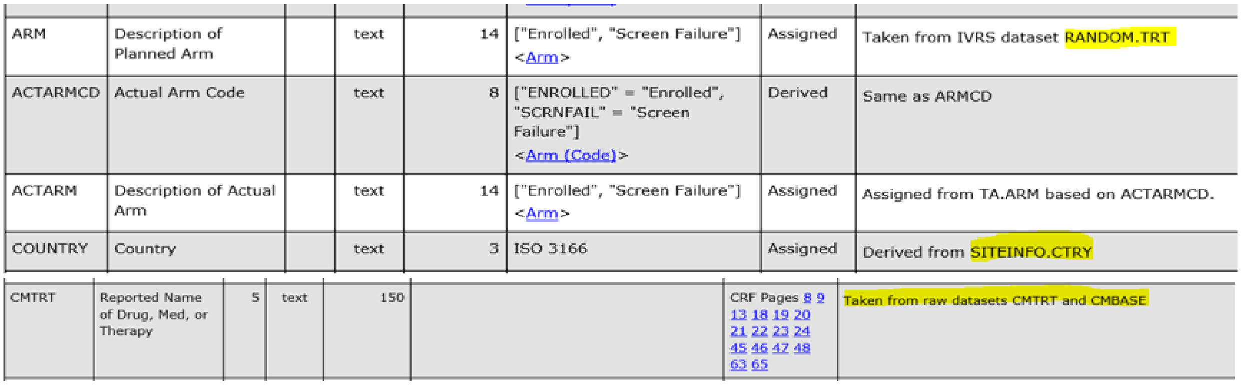
SDTM is considered the main data source of your data submission; as such, in the context of a data submission package, “it does not matter” how you migrated your legacy datasets to SDTM and which transformation you have made to your legacy data to make it compliant to SDTM standard. Figure 3 (next page) shows an example of improper SDTM define.xml specifications where the author of this define.xml included references to legacy datasets, while these datasets are not part of the submission. Of course migration specifications should be documented in some study documentation such as an excel file documenting how you did migrate your legacy datasets to SDTM.

> *Migration specifications should not be provided in the SDTM define.xml, as the reviewer will not understand what you are referring to*

Make sure your reviewer guide has relevant details

The reviewer guide (RG) has become a key document in every data submission package. This is the document that will drive the reviewer once he/she will receive the data submission package. For this reason you should not hesitate in repeating details that might be already mentioned in other documents since the “redundancy” of information reported in the RG is done on purpose. Often the RGs in the CDISC package I have reviewed contain either too many un-useful details or they were inversely lacking of important details.

[2] Note that while you might get an error from P21 while validating your define.xml, this is not an issue anymore when you submit to the FDA so in your define fields you can now have longer than 1000 characters (https://www.pinnacle21.com/forum/dd0086-maximum-length-1000-characters-data-attributes)

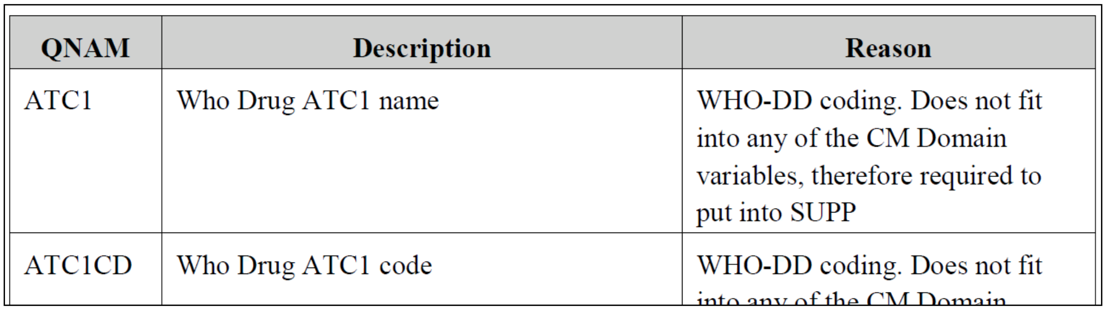


ARM	Description of Planned Arm		text	14	["Enrolled", "Screen Failure"] <Arm>	Assigned	Taken from IVRS dataset RANDOM.TRT
ACTARMCD	Actual Arm Code		text	8	["ENROLLED" = "Enrolled", "SCRNFAIL" = "Screen Failure"] <Arm (Code)>	Derived	Same as ARMCD
ACTARM	Description of Actual Arm		text	14	["Enrolled", "Screen Failure"] <Arm>	Assigned	Assigned from TA.ARM based on ACTARMCD.
COUNTRY	Country		text	3	ISO 3166	Assigned	Derived from SITEINFO.CTRY
CMTRT	Reported Name of Drug, Med, or Therapy	5	text	150		CRF Pages 8 9 13 18 19 20 21 22 23 24 45 46 47 48 63 65	Taken from raw datasets CMTRT and CMBASE

Figure 3: An SDTM define.xml referencing an legacy datasets not included in the submission package

Some examples:

- Among the standard sections of the Clinical Study Data RG (cSDRG), the SDTM reviewer guide, there is a standard section where all supplemental qualifiers are described for each domain included in the SDTM package. The standard template developed by PhUSE expects to report all QNAM and the associated description (QLABEL). It could be worthwhile to provide also a rationale in a separate column with the reason why the variable was not reported elsewhere, either in the parent domain or in another domain. The example in figure 4 is from the CM section; in there added supplemental qualifiers containing WHO-DD medical coding and therefore we have provided the reason why they were not reported in the parent CM domain
- On the contrary, some other cSDRGs were abusive with details with some data-management folks thinking the cSDRG is a data management report. You can of course provide relevant details concerning the quality of the data, but this should be limited to important items



QNAM	Description	Reason
ATC1	Who Drug ATC1 name	WHO-DD coding. Does not fit into any of the CM Domain variables, therefore required to put into SUPP
ATC1CD	Who Drug ATC1 code	WHO-DD coding. Does not fit into any of the CM Domain

Figure 4: A cSDRG containing details and rationale of information stored in the Supplemental Qualifier

- The conformance section of the RG should be not a simple list of issues detected by the validator tool; a clear rationale justifying the unresolved issues should be provided. "Resolvable" issues should be all addressed so they are no longer an issue. Figure 5 shows some examples of bad or insufficient rationale

> *The conformance section of the RG should be not a simple list of issues detected by the validator tool a clear rationale justifying the unresolved issues should be provided*

Rationale provided in the cSDRG	What is wrong with the justification?
Issue: “NULL value in AEDECOD variable marked as Required” Explanation: “Terms were not coded in the database”	Incomplete coding might be the object of a rejection to PMDA and major concern of the FDA. You need a strong rationale and therefore the explanation should have provided the reason why the term was not coded
Issue: “Missing FADY variable, when FADTC variable is present Explanation: “Variable not used”	Although –DY variable is permissible and sponsor could omit it, the FDA Study Technical Conformance Guide requires –DY variable to be included when –DTC is included in the data. Simply saying “Variable not used” does not matter!
Issue (AE): “Permissible variable with missing value for all Records” Explanation: “No data has been collected”	This is about seriousness criteria. It would have been better to clearly specify in the explanation for which variables this issue concern and probably either say that no serious AE with that specific criteria did occur (and in that in case you can also omit the criteria variable) or mention to the reviewer that the study CRF was not collecting such a detail (if that was the case)
Issue: “Invalid value for --TEST variable” Explanation: “Many instances of --TEST >40 characters. --TEST values are directly assigned from the labels taken from the Case Report Form to have clear understanding of the test code and therefore text was not changed.”	This is a wrong implementation! FCTEST should have been abbreviated and full text specified in the SDRG
Issue: “NULL value in SEX variable marked as Required” Explanation: “Data Issue: Sex is collected in the raw data”	Does it mean the sex was collected but for some subject the information was not available in the original data?
Issue: “Inconsistent value for Standard Units” Explanation: “Data Issue: We have not been able to convert these to standard units”	More details about laboratory parameter and unit concerned should have been mentioned

Figure 5: Providing a good rationale to unresolved issues

CLARIFICATIONS ABOUT SDTM IMPLEMENTATION

aCRF Clarifications

As of today there is not any industry standard on how the SDTM Annotated CRF (aCRF) should be created, but the CDISC “Metadata Submission Guideline (MSG) for SDTM IG” is a valid document sponsors and CROs should refer to with regards to aCRF, in particular section 4 “Guideline for Annotating and Bookmarking and CRFs”.

The aCRF is a critical document since it visually describes the content of your SDTM datasets; however, again form the packages I have reviewed, very often basic rules are not followed.

Domain Name Annotation

From the CDISC MSG “*Each domain that is represented on a CRF page should have its own annotation on the left side of the CRF page with the 2 letter domain code and domain name (e.g. IE = Inclusion/Exclusion)*”. Figure 6 provides a wrong interpretation of such a statement (left) and a correct interpretation (right).

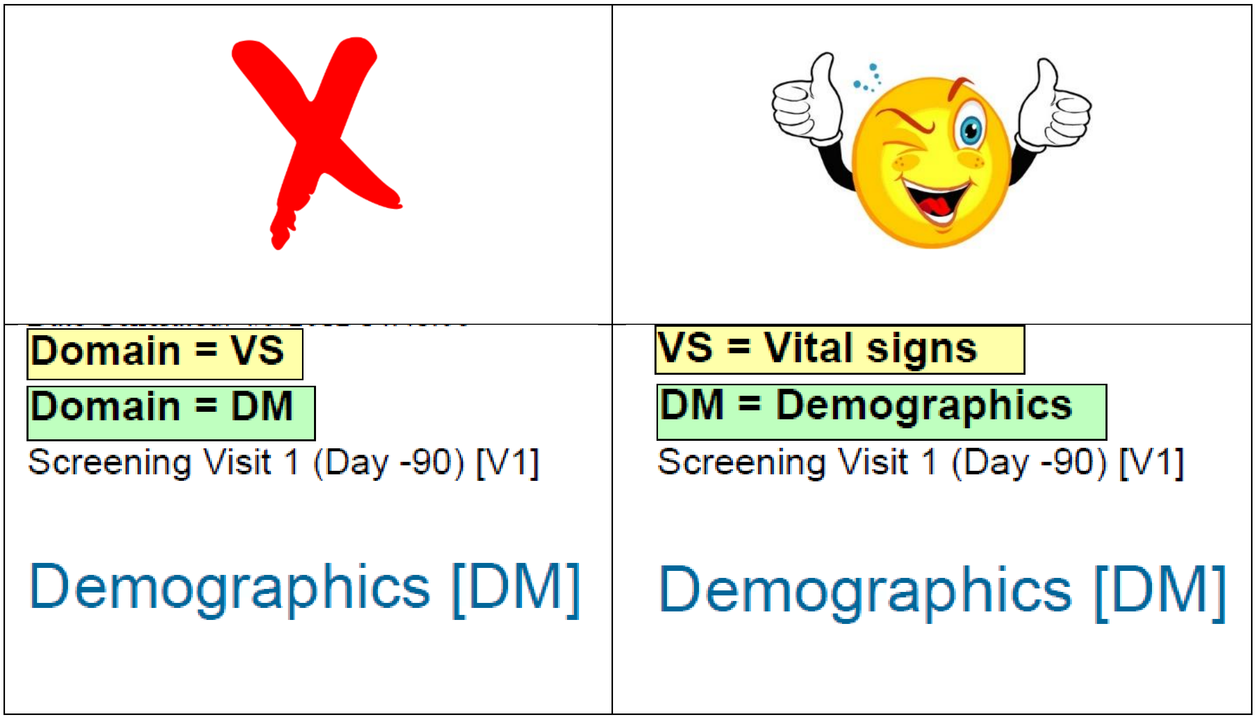


Figure 6: Wrong vs Correct Domain Name Annotation

Requirements for Bookmarking and Table of Contents (TOCs)

From the CDISC MSG “*Annotated CRFs included in the eCTD should be bookmarked 2 ways (dual bookmarking): bookmarks by time-points, often analogous to planned visits in the study, and bookmarks by CRF topics or forms. SDTM domains do not necessarily have a 1-to-1 relationship with CRF topics or forms, nor is the reverse true. For example, in the annotated CRF, both DM and SC are collected on the Demography panel, while SC data are collected from the Enrolment Form and the Demography pages*”.

Figure 7 is an example of an incorrect interpretation of this statement (left) and a correct interpretation (right). In the first example under “Domain” the sponsor has simply used the name of the original legacy data domain, while they should have used the name of the CRF form (the data being collected). In the second example ‘By domains’, instead of bookmarking the name of the applicable visits for the Clinical Events domains, they bookmarked the CRF page number, which of course does not make sense.

SDTM Controlled Terminolgy (CT) do not need to parallel the SDTM Ig version

CDISC SDTM Ig version does not need to parallel the CDISC CT version

One of our sponsor appointed a CRO specialized in Phase I studies. The CRO had a consolidated standardized system end-to-end from data collection to SDTM production to ADaM and TLFs[3] generation. However, the entire CRO ‘system’ was based on the SDTM Ig 3.1.3 and CDISC Controlled Terminology from more or less the same “era” of the SDTM Ig 3.1.3 (2013). We didn’t want to ask the CRO to update

[3] TLF = Tables Listings and Figures

their system for the sponsor study; this was also not worth given the fact for the type of study, phase I studies, the differences between SDTM Ig 3.1.3 and 3.2 were minimal. Nevertheless, we asked the CRO to at least update the controlled terminology to a more recent version to avoid any major differences when data will be eventually pooled with other studies. The CRO answer was as follows: "*This is XXX Inc. standard, the development of SDTM IG 3.1.3 has been done in 2013*". The CRO misinterpretation here is to think that the version of the SDTM Ig had to "parallel" the CDISC CT version, but there are actually no requirements to have Ig and CDISC CT from the same period and therefore the CRO, despite working with SDTM Ig 3.1.3, could have implemented a more recent version of the CDISC CT e.g. one of the 2018 CDISC CT.

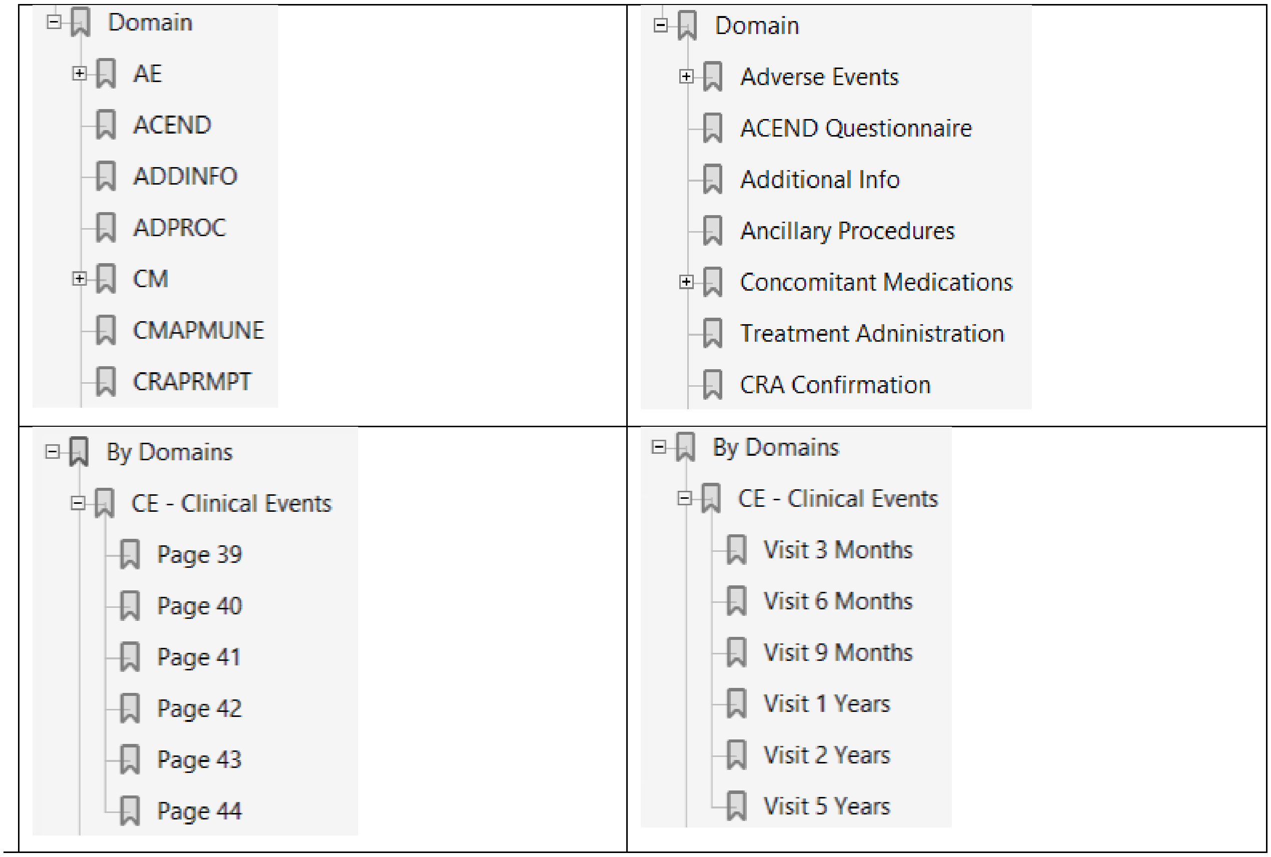


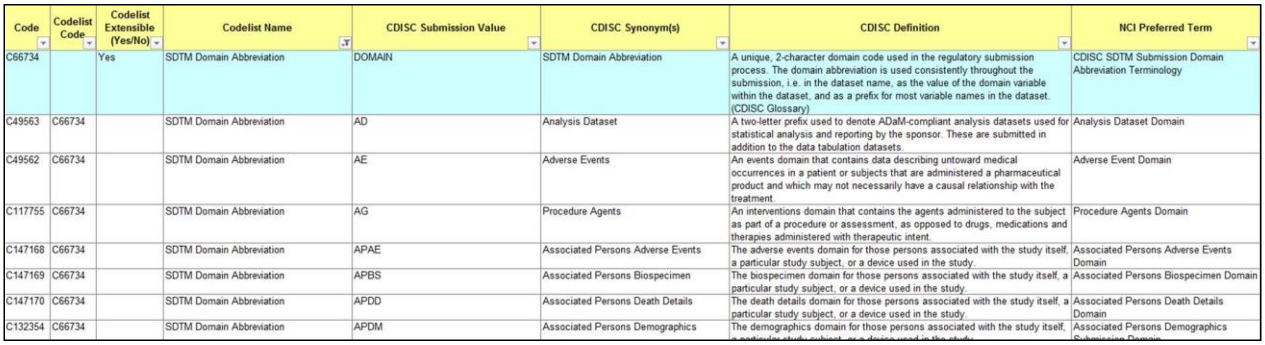
Figure 7: Wrong vs Correct SDTM aCRF Bookmarking

Standard terminology for SDTM domain names

Not all domains names are described in the latest SDTM Ig, but they are referenced in the latest CDISC CT version. The CDISC CT contains the list of "reserved" standard domain names up to the date the CDISC CT was released, like any other standard terminology (see figure 8 next page).

Do not change the "case" of a standard dictionary

Whenever data are coded with an external dictionary such as MedDRA, original "case" should be kept e.g. do not uppercase. If you are the "receiver" of the data from data management, make them aware of the issue.



Code	Codelist Code	Codelist Extensible (Yes/No)	Codelist Name	CDISC Submission Value	CDISC Synonym(s)	CDISC Definition	NCI Preferred Term
C66734		Yes	SDTM Domain Abbreviation	DOMAIN	SDTM Domain Abbreviation	A unique, 2-character domain code used in the regulatory submission process. The domain abbreviation is used consistently throughout the submission, i.e. in the dataset name, as the value of the domain variable within the dataset, and as a prefix for most variable names in the dataset. (CDISC Glossary)	CDISC SDTM Submission Domain Abbreviation Terminology
C49563	C66734		SDTM Domain Abbreviation	AD	Analysis Dataset	A two-letter prefix used to denote ADaM-compliant analysis datasets used for statistical analysis and reporting by the sponsor. These are submitted in addition to the data tabulation datasets.	Analysis Dataset Domain
C49562	C66734		SDTM Domain Abbreviation	AE	Adverse Events	An events domain that contains data describing untoward medical occurrences in a patient or subjects that are administered a pharmaceutical product and which may not necessarily have a causal relationship with the treatment.	Adverse Event Domain
C117755	C66734		SDTM Domain Abbreviation	AG	Procedure Agents	An interventions domain that contains the agents administered to the subject as part of a procedure or assessment, as opposed to drugs, medications and therapies administered with therapeutic intent.	Procedure Agents Domain
C147168	C66734		SDTM Domain Abbreviation	APAE	Associated Persons Adverse Events	The adverse events domain for those persons associated with the study itself, a particular study subject, or a device used in the study.	Associated Persons Adverse Events Domain
C147169	C66734		SDTM Domain Abbreviation	APBS	Associated Persons Biospecimen	The biospecimen domain for those persons associated with the study itself, a particular study subject, or a device used in the study.	Associated Persons Biospecimen Domain
C147170	C66734		SDTM Domain Abbreviation	APDD	Associated Persons Death Details	The death details domain for those persons associated with the study itself, a particular study subject, or a device used in the study.	Associated Persons Death Details Domain
C132354	C66734		SDTM Domain Abbreviation	APDM	Associated Persons Demographics	The demographics domain for those persons associated with the study itself,	Associated Persons Demographics

Figure 8: CDISC CT Study Domain Abbreviation

Apply Standard CDISC CT whenever you can

Very often people just stick to the SDTM Ig or think that a supplemental qualifier is just a “trash bin” where we can put items we don’t know where else to put. The fact that we stored something in supplemental qualifier dataset does not mean you don’t have to be accurate as much as in the parent domain when migrating data from legacy datasets. For example, terminology such as YN or ISO formats, such as the one for date, they should be applied whenever it is possible. Some examples:

- Is your Supplemental Qualifier a date? ISO format should be applied and therefore original date should be converted to the YYYY-MM-DD ISO format (assuming here it is only containing the date part)
- Does your Supplemental Qualifier have only Yes/No results? Then convert “Yes” to “Y” and “No” to “N”
- Does your result for a specific parameter in a finding dataset have only Yes/No results? Then convert “Yes” to “Y” and “No” to “N”. Similarly for Negative/Positive

SUPP is not a simple “Trash Can”, keep it clean!!!

Make use of subset CT in your define.xml

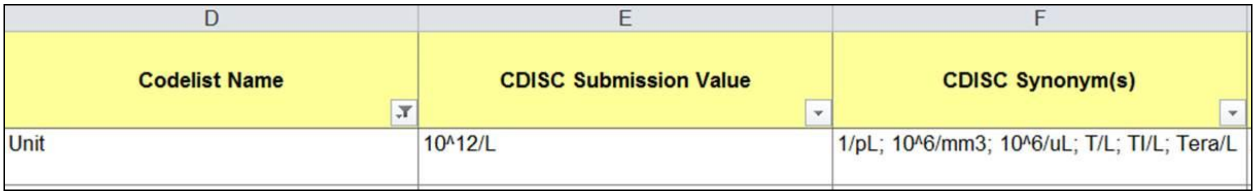
Often I have seen define.xml with poor metadata handling. One example is to simply include the entire CDISC CT even if some items of the CDISC CT do not apply at all to that specific variable. The following are just some recommendations on how to properly document CDISC CT in define.xml:

- There are terminologies with broad scope e.g. UNIT has unit applicable to laboratory data, medication units, etc. In define.xml you can have subset of units e.g. LBUNIT, CMUNIT, etc. where only applicable units are listed in each subset
- Only terms that occurred or CRF pre-printed terms should be reported in the CT definition in the define.xml e.g. if from the CRF 5 races are expected but you had only WHITE subjects, all 5 expected races should be part of the CT
- Do not include in the CT definition all terms if all are not applicable e.g. do not include all terms from the UNIT CDISC CT (>500 terms)
- If a subset of CT is defined in define.xml, you still need to reference “Codelist Code” and “Code” as per CDISC CT. For example, if you define a CT CMUNIT being a CT subset of the standard CT UNIT, the “Codelist Code” and “Code” of the term coming from the CT UNIT should be referenced in your define.xml

Column E (CDISC Submission Value) is the one to be ‘used’ in the dataset

Once again here my French colleagues might say “ca va sans dire” but column E from the CDISCT CT excel file (see figure 9) is the one to be used in the submitted SDTM dataset. Column F, CDISC Synonym(s), can be not used. Column F is there to facilitate the conversion to the appropriate CDISC term (CDISC Submission Value).

For example from figure 9 the CDISC Submission Value for the unit is ‘10^12/L’ as in column E, while the column F values are there to facilitate the unit conversion in case your data contains similar terms e.g. 1/pL, 10^6/mm3, 10^6/uL, T/L, TI/L and Tera/L, are same as 10^12/L, so in your SDTM dataset you will need to use this term and not the one in column F.



D	E	F
Codelist Name	**CDISC Submission Value**	**CDISC Synonym(s)**
Unit	10^12/L	1/pL; 10^6/mm3; 10^6/uL; T/L; TI/L; Tera/L

Figure 9: CDISC CT Submission Value vs CDISC Synonyms

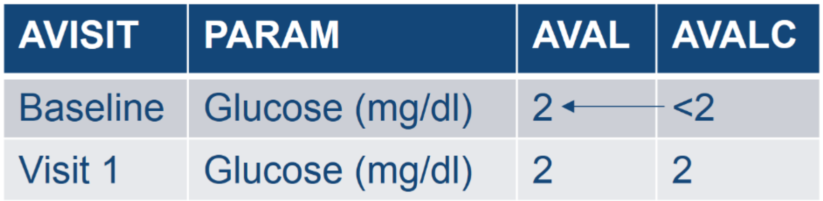
CLARIFICATIONS ABOUT ADAM IMPLEMENTATION

Improper use of AVAL/AVALC in ADaM BDS

The following is a request for help from a colleague seeking advice on how to solve or justify an issue raised by the validator e.g. P21 community.

> *I’m getting the following error message from P21 “Inconsistent value for AVALC”. This is because in my study for one parameter I have values such as ‘<2’ (in AVALC) that have been imputed to a numeric value of ‘2’ (in AVAL). This was done according to the Statistical Analysis Plan. However I have also numeric results reported exactly as ‘2’ and I get this error even if I set my AVALC equal to Null when my result does not contain any sign*

In this ADaM dataset some sort of imputations were applied so that a value below a certain level of quantity was assigned a numeric ‘finite’ value e.g. <2 equal to 2. This was generating an issue given the fact there were also some observed ‘finite’ result equal to 2 (see figure 10).



AVISIT	PARAM	AVAL	AVALC
Baseline	Glucose (mg/dl)	2 ←	<2
Visit 1	Glucose (mg/dl)	2	2

Figure 10: Missing one-to-one match between AVAL and AVALC

This is going against one ADaM BDS rule requiring, for a given parameter, a one-to-one relationship between the variables AVAL and AVALC when both populated. And here for the same parameter for a value in AVAL, the value of “2”, two distinct values occur in AVALC, the value of “2” and the value of “<2”

The issue can be avoided by removing the AVALC variable given the fact there is no need to have both AVAL and AVALC with non-missing values for a given parameter. My recommendation is to always use AVAL when the result is of numeric 'nature', otherwise use AVALC if the result is of 'character' nature, but not both. In the above issue I consider the parameter of numeric nature as the value <2 is still representing a numeric result. In this case you don't need AVALC and eventually for traceability purpose you can still keep either --ORRES or --STRESC from the source SDTM domain.

My recommendation is to always use AVAL when the result is of numeric 'nature', otherwise use AVALC if the result is of 'character' nature, but not both

The bottom line is that in most of the situations for a given parameter in a BDS ADaM, only AVAL or AVALC will be not null. The only situation in which you may have both AVAL and AVALC not null is with a results of character nature, such as Adverse Event Severity collected as Mild/Moderate/Severe (AVALC) where for analysis purposes e.g. to have a way of ordering the severity, you assign a numeric result for example 1 for Mild 2 for Moderate and 3 for Severe (AVAL).

Where is my Traceability

One of our sponsor requested to review a CDISC ADaM package developed internally by their Biostatistics team. The focus was especially on ADaM where the sponsor did not have that much experience.

In my Issues log containing all issues or recommendations I gave to the sponsor to help them being conformant and enough "traceable", I was concerned about the way one of their ADaM dataset supporting the analysis of ECG data was built. The sponsor team derived the mean of the three ECG measurements (triplicates); up to here nothing wrong – this was correctly derived in ADaM. However, I recommended keeping also the three original records from which the derived parameter was created. The answer from the sponsor was a bit "unsympathetic": "*The original records were not retained in ADEG because they are in SDTM.EG"*. Although there is no 'obligations' from the ADaM Ig to keep all records from SDTM when creating ADaM datasets, it is a good attitude to keep these records when these records are the source of records (parameters) derived in ADAM.

Among the four fundamental ADaM principles, traceability is in my opinion (but not only my opinion) the most important. From the ADaM Ig traceability is defined as follows:

> "*The property that enables the understanding of the data's lineage and/or the relationship between an element and its predecessor(s). Traceability facilitates transparency, which is an essential component in building confidence in a result or conclusion. Ultimately, traceability in ADaM permits the understanding of the relationship between the analysis results, the ADaM datasets, the SDTM datasets, and the data collection instrument. Traceability is built by clearly establishing the path between an element and its immediate predecessor. The full path is traced by going from one element to its predecessors, then on to their predecessors, and so on, back to the SDTM datasets, and ultimately to the data collection instrument*".

Traceability in ADaM Ig is continuously "stressed" making this principle key for a good ADaM dataset. One aspect stressed in the ADaM Ig is the transparency that you provide by making your ADaM fully traceable:

> *"Retaining in one dataset all of the observed and derived rows for the analysis parameter provides the clearest traceability in the most flexible manner within the standard BDS. The resulting dataset also provides the most flexibility for testing the robustness of an analysis"*

So unless, you end-up with a huge ADaM dataset, keep original source records in your dataset: this will be beneficial for the traceability of your ADaM derived variables and ultimately it will be transparent for the reviewer which records have been used (and eventually which records have been not used or not *analysed).*

My PARAM should be enough and the concept of analysis-ready

Again, while trying to assist our sponsors and advising them, I was questioning the presence of a variable AVALU in most of their BDS-type ADaM datasets since I did not understand the need and their answer was "*We need it to store the unit of PARAM*". This is what I call a "decorative variable" since PARAM should already contain this information; from ADaM Ig "*PARAM must include all descriptive and qualifying information relevant to the analysis purpose of the parameter*" (ADAM Ig 1.1 Section 3.3.4) e.g. unit PARAM="Weight (kg)".

The other common "mistake", although not really a conformance issue, it is to create PARAM by concatenating for example all SDTM findings qualifiers of the --TEST variable, for example for ECG "Systolic Blood Pressure (mmHg), Sitting Position". This is not really needed unless you are doing an ECG analysis based on the subject position. Again this is clear from the ADaM Ig section 3.3.4: "*PARAM must include all descriptive and qualifying information relevant to the analysis purpose of the parameter*", this means, in few words, that in most of the cases concatenating the parameter name and its unit should be enough (this is what usually goes into the statistical output).

This satisfies one of the ADaM fundamental principle that requires your ADaM dataset to be "analysis-ready". This concept was, and still is, the subject of some misunderstanding. It is often (mis-)understood as "one-proc-away" meaning that we cannot make for example any merge with any other datasets prior to call for example the SAS Statistical procedure, but the requirement of being analysis-ready "simply" refers to *a particular data point/result on a table*, and any given statistical table might use multiple analysis datasets and multiple statistical procedures.

"SUPERFLUO" ADaM SUBMISSION

How many analysis ADaM datasets do I need to create given the fact my study SDTM has 22 subject datasets?

There is of course no correct answer without knowing the details of the statistical analysis to be performed. ADaM, as opposed to SDTM, is analysis-driven, meaning that you do not need to create an ADaM dataset for laboratory data if your Statistical Analysis Plan does not plan for any aggregate analysis on laboratory data (e.g. you might create listings directly from your SDTM if you don't need any major derivation) and very likely you will not need an ADIE ADaM dataset. *ADIE what? An ADaM dataset for violated Inclusion and Exclusion Criteria.* This is really not needed!

These are some of the recommendations we give on this topic during the CDISC ADaM training:

<u>What analysis datasets are required for a submission?</u>

ADaM ADSL dataset is required. FDA expects additional ADaM datasets to support primary and secondary analysis. From the FDA Technical Conformance Guidance: "*Sponsors should submit ADaM dataset to support key efficacy and safety analysis and …. the primary and secondary endpoints of a trial … should be provided as well*".

<u>Should there be an ADaM dataset for every SDTM domain?</u>

No. There is no requirement that every SDTM domain has a corresponding analysis dataset. ADaM datasets are needed to support the Statistical Analysis Plan, and they are not simply created as a mechanistic duplication of SDTM data.

CONCLUSION

One of my wife's favourite TV-shows is 'Quattro Ristoranti' (Four Restaurants). In each episode of the show, four restaurants of the same style are assessed and the one getting the best evaluation wins the prize. One of the first things the TV presenter Alessandro Borghese, a famous Italian chef, does while visiting the restaurant is to assess (of course!) the kitchen and how much the kitchen and its tools are cleaned. This assessment could have a big impact on the final outcome regardless of the quality of the food served in the restaurant .... the state of the kitchen and its cleanliness influences Borghese's faith in the chef's work.

This is exactly what could happen in a data submission to health authorities such as the FDA: *the efficacy and safety of your drug are of course what matter, but lack of traceability, or poor or insufficient documentation might trigger questions and concerns from the reviewer*. While this might not impact the overall final outcome of your submission, approval could be delayed if the reviewer starts questioning what you have done by requesting changes, or new deliverables to clarify the aspects that were not sufficiently clear in your original submission.

> *The efficacy and safety of your drug are of course what matter, but lack of traceability, or poor or insufficient documentation might trigger questions and concerns from the reviewer*

When working with CROs it is important that sponsors clearly define what they expect as deliverables from them, not only which files are to be included in the CDISC package but also any position they might have on a particular aspect of the standard that give more than one option. Such requirement can be specified in a sponsor implementation guidance (complex) or in a simple checklist that could be used during the review; this could also be provided to the CRO so that the CRO is aware of the additional checks the sponsor will go through once the package is delivered.

As previously mentioned, the packages I have reviewed demonstrated how differently individual "users" and companies have approached the same mapping issue in SDTM or the same analysis "modelling" in ADaM; hopefully this will be reduced in future, hopefully as soon as the industry, eventually with the HA, will address all standard “nuances” and agree a common interpretation.

> *Don’t get bored, be patient, love and cure the standards! Have a passion for details as they might matter when you submit your data to an agency. This is what I try to pass on to my colleagues almost every day, the passion for the data especially when they are organized in a standard way*

REFERENCES

PhUSE Study Data Reviewer Guide (SDRG) and Analysis Data Reviewer Guide (ADRG) template. https://www.phuse.eu/css-deliverables

CDISC ADaM Implementation Guidance v1.1. https://www.cdisc.org/standards/foundational/adam

Metadata Submission Guideline (MSG) for SDTMIG. https://www.cdisc.org/standards/foundational/sdtmig

PhUSE Standard Analyses & Code Sharing – White Papers. https://www.phuse.eu/white-papers

FDA Study Data Technical Conformance Guide. https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm2005545.htm#guides

Portable Document Format (PDF) Technical Specifications Document. https://www.fda.gov/downloads/Drugs/UCM163565.pdf

FDA Data Standards Catalog.https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM340684.xlsx

FDA Technical Rejection Criteria for Study Data. https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM523539.pdf

RECOMMENDED READING

About Traceability

- S. Minjoe, T. Petrowitsch. 2014. “Traceability: Plan Ahead for Future Needs” PhUSE-EU, London
- Traceability and Data Flow PhUSE-CSS WG. http://www.phusewiki.org/wiki/index.php?title=Traceability_and_Data_Flow
- Summary of Traceability References PhUSE-CSS Wiki Page. http://www.phusewiki.org/wiki/index.php?title=Summary_of_Traceability_References
- Angelo Tinazzi. “Lost in Traceability, from SDTM to ADaM …. finally Analysis Results Metadata”. 2016. CDISC EU Interchange. Barcelona

About Data Submission and Proper Documentation

- Angelo Tinazzi, Cedric Marchand. “An FDA Submission Experience Using the CDISC Standards” 2016. PhUSE. Barcelona
- Varum Debbeti. “How to Prepare High-quality Metadata for Submission”. 2018. PhUSE US Connect. Raleigh
- D. Roulstone. “Do's and Don'ts of Define.xml”. PhUSE-EU. 2018. Frankfurt
- Markus. Stoll, Laura. Phelan, Angelo Tinazzi. “A SDTM Legacy Data Conversion”. PhUSE-EU. 2018. Frankfurt

Other Presentations

- C. Paul, S. Sturm. “Guideline for submission ready aCRF”. 2018. PhUSE-EU. Frankfurt

CONTACT INFORMATION

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

Angelo Tinazzi
Cytel Inc
Angelo.tinazzi@cytel.com
www.cytel.com
https://www.cytel.com/blog/topic/statistical-programming check for my blog series “The Good Data Submission Doctor”

Any brand and product names are trademarks of their respective companies.