PhUSE 2014

Paper CD01

Challenges with the interpretation of CDISC

-

Who can we trust?

Linda Palm Simonsson, I-Mind, Lund, Sweden

ABSTRACT

Many smaller companies have none or very little knowledge of biometrics and turns to one or several CROs to help them conduct their trials or a submission. A small company expects the CRO to be the expert and since many small companies lacks a biometric department they also lack internal resources, routines and expertise to do a qualified QC of the SDTM and the ADaM packages received from a CRO. Sometimes the difficulties for the sponsor begins already at understanding what's included in the proposal vs what's needed according to FDA or other agency. Other difficulties may be inconsistency between e.g. SDTM IG and FDAs requirements.

This presentation will address some errors, misunderstandings and difficulties seen in the interpretation of different documents and showing some inconsistencies between different sources.

INTRODUCTION

Over the years of working with CDISC, both me and my colleagues has encountered lots of different interpretations of the implementation of different CDISC standards, this paper shows some of the more common errors or inconsistencies found and some of my personal thoughts about what should have been done instead.

EMPTY DATASETS

All documentation seems to agree that empty datasets should not be submitted. But how should it be documented that you didn't submit an expected dataset?

> **SDTM IG – Section 3.2**
> In the event that no records are present in a dataset, the empty dataset should not be submitted and **should not be described in the define.xml** document. The annotated CRF will show the data that would have been submitted had data been received; **it need not be re-annotated to indicate that no records exist.**

While the FDAs website says

> **FDA - Supplemental Information for Planning a CDISC Formatted Submission**
> If no data is collected for a specific domain, **annotate on the CRF and the define.xml** but do NOT submit an empty dataset (Update 8/12/10)

The question that probably pops up is "should I or shouldn't I include empty datasets in the define.xml"? Searching the web for this I found (some years ago) that you should have it in define.xml (at least for IE since IE is a required domain) with a description like "no subject has satisfied this criteria" otherwise Janus (FDAs warehouse) would not validate the define.xml correctly. Although it seems most common not include empty datasets in define.xml.

In 2013 FDA and PhUSE published a template for a Study Data Reviewer's Guide (SDRG), where all planned but not submitted dataset should be documented.

Were any domains planned, but not submitted because no data were collected?

Yes

If yes, list domains not submitted:

IE – All subjects met inclusion/exclusion criteria.

SUPPDM – For all subjects, race was one of those pre-specified on the CRF.

Specification for "other" race was not needed.

So the right (and only) place to document the empty datasets is in the SDRG.

ARM

The variables ARM and ARMCD have a very central role in SDTM, and the SDTM IG has already a rule for how screening failures should be handled. But in February 2014 FDA released a new draft guideline (which aren't for implementation – Yet (!)) that contradicts the SDTM IG. So please watch out for the final version before you start to make any changes.

SDTM IG v3.2 (and previous versions)

Data for screen failure subjects, if submitted, should be included in the Demographics dataset, with ARMCD = "SCRNFAIL" and ARM = "Screen Failure".

STUDY DATA TECHNICAL CONFORMANCE GUIDE Technical Specifications Document (Draft feb 2014)

Screen failures, when provided, should be included as a record in DM with the ARM field left blank.

TRIAL DESIGN

The Trial Design datasets has for many years been more or less ignored, since they have no patient data. Different vendors have different approaches to if they include all or some trial design domains. Lots of proposals/contracts does not mention them at all (so you can't know if it's included or not) while others clearly states that TS is out of scope. There are also CROs that deliver everything but the TS. But what do we actually need to include?

FDA - Supplemental Information for Planning a CDISC Formatted Submission

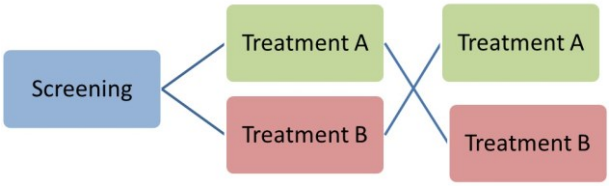
Trial Design datasets provide a standard way to describe the planned conduct of a clinical trial**. At a minimum, the Trial Summary (TS) domain should be submitted whenever possible.** Additional variables not listed in Appendix C3 of the SDTM IG v3.1.2 may be added as long as they are explained in the define.xml (Update 12/22/10)

Also newer versions of OpenCDISC will give an error message whenever TS is not included.

So in conclusion we should at least include the TS dataset in a submission.

SUBJECT ELEMENTS

One sponsor needed to convert a legacy cross-over study to SDTM for a submission and the proposal from the CRO stated that *"SE is out of scope"*.



FDA - Supplemental Information for Planning a CDISC Formatted Submission
Subject Elements (SE) and Subject Visits (SV) describe actual elements and visits for each subject. These datasets **should be submitted whenever possible**. This is **especially important when the trial includes multiple treatment periods**.

Since FDA expects to see those datasets, we should whenever possible provide SE and SV.

CREATING A NEW DOMAIN

There are very strict rules for when we can create a new domain or not, but these are not always followed.

SDTM IG – Section 2.6
Confirm that none of the existing published domains will fit the need. A custom domain may only be created if the data are different in nature and do not fit into an existing published domain. Group and separate data within the domain using --CAT, --SCAT, --METHOD, --SPEC, --LOC, etc. as appropriate.

Look for an existing, relevant domain model to serve as a prototype. If no existing model seems appropriate, choose the general observation class (Interventions, Events, or Findings) that best fits the data by considering the topic of the observation.

CDER Common Data Standards Issues Document
Both the SDTM and SEND Implementation Guides allow for the creation of custom domains if the data do not fit into an existing domain. **Prior to creating a custom domain, sponsors should confirm that the data do not fit an existing domain and also check the CDISC website for domains added after the most recent published implementation guide.** If necessary, sponsors should follow the recommendations in the implementation guides for how to create a custom domain.

Below is three examples of when new domains shouldn't have been created

Example #1: Created a custom made finding domain for "Cardiac History"

Cardiac History

- High blood pressure? ☐ Yes ☐ No
- Chest pain on exertion? ☐ Yes ☐ No
- Previous heart attack? ☐ Yes ☐ No ___/___/_____
  dd mmm yyyy

In this chase the person creating the SDTM mistook "Cardiac History" for data that will into a Finding domain and didn't recognize it as Medical History of special interest. The programmer should have stored it in Medical History

(MH) with MHCAT="Cardiac History", while the ordinary Medical History could have had MHCAT="Relevant Medical History". Variables like MHTERM, MHPRESP, MHOCCUR and MHSTDTC should be used to present the data.

Example #2: A CRO followed IG 3.1.2 and had to create SDTM for an Oncology study.
(At this time point IG 3.1.3 was already finalized)

- They created a custom made domain starting with Y for the tumor result data

In this case the programmer/SDTM mapper should have checked the CDISC website for domains added after IG 3.1.2, since TU, TR and RS were included in SDTM IG 3.1.3.

If a new domain has been defined in a later version (even if the later version is only in draft), you must use the later version instead of creating your own.

Example #3: A third CRO acknowledged that they had clinical events, but since they already used CE for other clinical event data they created a new domain, XX.

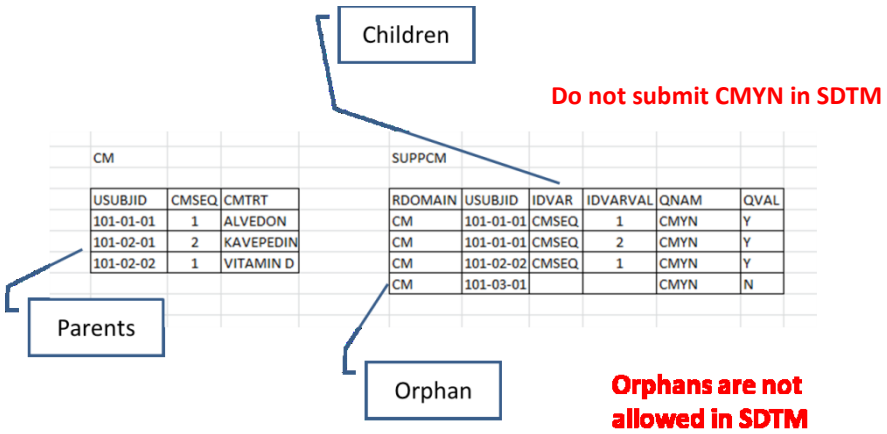
All clinical event data should have been in CE, separated by the CECAT variable

ORPHANS

An orphan is an observation in e.g. SUPP-- or RELREC that has no related observation in the domain referred to in RDOMAIN (the parent domain).

I have seen orphans in several mapping specification, and the following one is one of the most typical. Many programmers or data managers seems to try to map everything on the CRF instead of annotate, e.g. cleaning variables, with "not submitted".

The CDASH-variable "Any Concomitant Medications (Y/N)?" was mapped to be in SUPPCM, so every time they save the answer "No" in SUPPCM, they created an orphan. All entries in a SUPP-- must have linked data in the parent domain



.

SPONSOR-DEFINED REFERENCE PERIOD

Another topic that seems a bit hard is the sponsor-defined reference period in comparison with variables ending in --STRF and --ENRF.

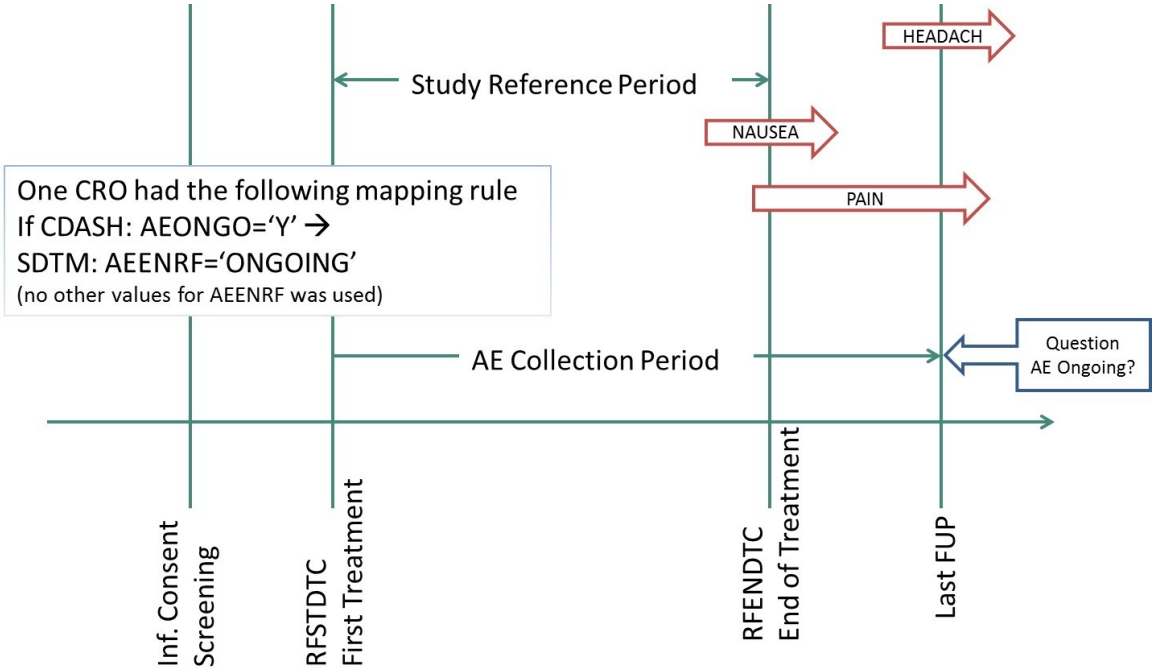
SDTM IG – Section 4.1.4.7

--STRF is used to identify the start of an observation relative to the sponsor-defined reference period.
--ENRF is used to identify the end of an observation relative to the sponsor-defined reference period.

Allowable values for --STRF and --ENRF are “BEFORE”, “DURING”, “DURING/AFTER”, “AFTER”, “COINCIDENT”, “ONGOING”, and “U” (for unknown).

As an example, a CRF checkbox that identifies concomitant medication use that began prior to the study treatment period would translate into CMSTRF = “BEFORE” if selected; similarly, a CRF checkbox that identifies concomitant medication use that continues after the study treatment period would translate into CMENRF = “ONGOING” if selected.

A typical example:
The study reference period was in one study defined as the period from first to last exposure of the study treatment. After the last treatment the subjects are followed for three more weeks. Adverse events are collected from the first exposure to treatment to the last follow-up (three weeks) after the reference period ends. Adverse event still ongoing at the last follow-up were checked as ongoing in the CRF (AEONGO=Y).



Since AE ongoing isn’t asked at the end of the Study Reference Period we cannot use the AEENRF variable, instead we should use AEENRTPT and AEENTPT, e.g.

- AEENTPT=‘Last Follow-Up’
- AEENRTPT= ‘BEFORE’ (if AEONGO=‘N’ and AENEDTC<Last Follow-Up)
  ‘COINCIDENT’ (if AEONGO=‘N’ and AENEDTC=Last Follow-Up)
  ‘ONGOING’ (if AEONGO=‘Y’)

USE OF --PRESP WITHOUT –OCCUR

SDTM IG – Section 4.1.5.7

The --PRESP variable is used to indicate whether a specific intervention (--TRT) or event (--TERM) was solicited. The --PRESP variable has controlled terminology of Y (for "Yes") or a null value. It is a permissible variable, and **should only be used when the topic variable values come from a pre-specified list**.

The --OCCUR variable is used to indicate whether a pre-specified intervention or event occurred or did not occur. It has controlled terminology of Y and N (for "Yes"ll and "No"). It is a permissible variable and **may be omitted from the dataset if no topic-variable values were pre-specified**.

One CRO used the pre-specified variable in Substance Use without specifying if it occurred or not in SU, instead the saved the answer as collected in SUPPSU.

USUBJID	SUTRT	SUPRESP
101-01	CIGARETTES	Y
101-01	PIPE	Y
101-02	CIGARETTES	Y
101-02	PIPE	Y
101-03	CIGARETTES	Y
101-03	PIPE	Y

SUPPSU.QNAM	SUPPSU.QVAL
USER	No
USER	Current User
USER	Current User
USER	No
USER	Ex-Smoker
USER	No

The use of SUPPSU is good, since they probably like to use the variable later in their ADaM dataset, but looking at the examples in the implementation guide for SU: SUENTPT and SUENRTPT together with SOCCUR should have been derived, like:

```
if USER in ('Current User', 'Ex-Smoker') then do;
    SUOCCUR='Y';
    SUENTPT='SCREENING';
    if USER='Current User' then SUENRTPT='ONGOING';
    else if USER='Ex-Smoker' then SUENRTPT='BEFORE';
end;
else if USER in ('No') then SUOCCUR='N';
```

USUBJID	SUTRT	SUPRESP
101-01	CIGARETTES	Y
101-01	PIPE	Y
101-02	CIGARETTES	Y
101-02	PIPE	Y
101-03	CIGARETTES	Y
101-03	PIPE	Y

SUPPSU.QNAM	SUPPSU.QVAL
USER	No
USER	Current User
USER	Current User
USER	No
USER	Ex-Smoker
USER	No

SUOCCUR	SUENTPT	SUENRTPT
N		
Y	SCREENING	ONGOING
Y	SCREENING	ONGOING
N		
Y	SCREENING	BEFORE
N		

DERIVE OR IMPUTE

SDTM IG

Data stored in SDTM datasets include both raw (as originally collected) and derived values (e.g., converted into standard units, or computed on the basis of multiple values, such as an average).

When --ORRES is populated, --STRESC must also be populated, regardless of whether the data values are character or numeric.

AGE (Expected): Age expressed in AGEU. May be derived from RFSTDTC and BRTHDTC.

VSSTRESC (Expected): Contains the result value for all findings, copied or derived from VSORRES **in a standard format or standard units.**

CDER Common Data Standards Issues Document

For a given test, all values of --STRESU should be the same. In some cases --TESTCD may not be sufficient to uniquely identify a test.

SDTM should not include any imputed data. If there is a need for data imputation, this should occur in an analysis dataset, and the relevant supporting documentation to explain the imputation methods must be provided.

One CRO left AGE, --STRESC and –STRESN blank with the motivation that all derivations/imputations should be left for ADaM. After asking them to populate the variables they derived the age but for e.g. the temperature they just copied VSORRES into VSSTRESC without converting from F to C.

What we can conclude from this is that we are allowed to derive but not to impute! So what is the difference between deriving and imputing?

Imputing involves guessing, while deriving doesn't.

Derive

- You derive the temperature from F to C like C=(F-32)*5/9

Impute

- When we only have partial dates we may need to impute the date with rules described in the Statistical Analysis Plan

STANDARD QUESTIONNAIRES

Another common error is to handle standard questionnaires like this, with decoded values in both QSORRES and QSSTRESC:

QSTEST	QSORRES	QSSTRESC	QSSTRESN
Global Improvement	No change	No change	4
Global Improvement	Much Improved	Much Improved	2
Global Improvement	Minimally Improved	Minimally Improved	3

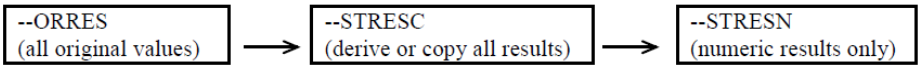
When the correct way for standard questionnaires is to handle it like this:

QSTEST	QSORRES	QSSTRESC	QSSTRESN
Global Improvement	No change	4	4
Global Improvement	Much Improved	2	2
Global Improvement	Minimally Improved	3	3

--STRESN should only be populated if --STRESC contains a number (and only a number)

SDTM IG - Section 4.1.5.1.1

When --ORRES is populated, --STRESC must also be populated, regardless of whether the data values are character or numeric. The variable, --STRESC, is derived either by the conversion of values in --ORRES to values with standard units, or by the assignment of the value of --ORRES (as in the PE Domain, where --STRESC could contain a dictionary-derived term). **A further step is necessary when --STRESC contains numeric values. These are converted to numeric type and written to --STRESN**. Because --STRESC may contain a mixture of numeric and character values, --STRESN may contain null values, as shown in the flowchart below.



When the original measurement or finding is a selection from a defined codelist**, in general, the --ORRES and --STRESC variables contain results in decoded format**, that is, the textual interpretation of whichever code was selected from the codelist.

In some cases **where the code values in the codelist are statistically meaningful standardized values or scores,** which are defined by sponsors or by valid methodologies **such as SF36 questionnaires, the --ORRES variables will contain the decoded format, whereas, the --STRESC variables as well as the --STRESN variables will contain the standardized values or scores**.

SDTM → ADaM TRACEABILITY

One ADaM mapping spec created ADSL.RACE in Title Case like:

```
If RACE='OTHER' then
    RACE=propcase(strip(RACE)||': '||SUPPDM.QVAL )
    (when QNAM='RACEOTH')

Else RACE=propcase(RACE);
```

ADaM IG v1.0

Any ADaM variable whose name is the same as an SDTM variable must be a copy of the SDTM variable, and its label, meaning, and values must not be modified. ADaM adheres to a principle of harmonization known as "same name, same meaning, same values."

The race of the subject is a required variable in ADSL. If the variable is not a copy of DM.RACE, then an additional differently named variable must be added.

This means that if a variable in ADaM has the same name as a variable in SDTM, it must be a direct copy of the SDTM variable. (In the next ADaM IG (v1.1) you will actually be allowed to shrink the length, but you are still not allowed to change anything else).

WHO CAN WE TRUST AND WHAT COULD A SPONSOR EXCPECT?

Looking at different documents, websites, vendors, programmers etc. we may wonder who or what we really can trust and people not so experienced in SDTM may think it's a jungle. We must trust CDISC, FDA etc., and we should rely on software's like OpenCDISC (but at the same time keep in mind that there are a few bugs, so review and examine every issue carefully). For the relationship between sponsors and CROs it's probably a learning curve, were they need to work together for several projects.

Sponsors should expect

- That all required and highly recommended items are included in the proposal (if not, it should be clearly stated).
- That the programming team have at least one members that could be considered senior when it comes to CDISC knowledge
- OpenCDISC or equivalent should be used for validation on both define.xml and SDTM/ADaM datasets for all drafts and final

CONCLUSION

Don't be afraid to ask your vendor to give you a more detailed proposal of what's included and if they have omitted anything that is required or recommended by e.g. FDA. Every company needs at least one in-house or contracted CDISC specialist. It would be good if we could have a common place where we linked all information together (PhUSE WIKI??)

REFERENCES

1. CDISC: SDTM IG 3.1.2, 3.1.3 and 3.2
2. CDISC: SDTM Terminology 2013-12-20
3. CDISC: ADaM IG 2.1
4. FDA: Study Data Specifications (SDS) v2.0
5. FDA: CDER Common Data Standards Issues Document v1.1
6. FDA: Supplemental Information for Planning a CDISC Formatted Submission
7. FDA: Study Data Technical Conformance Guide (Feb 2014 – Only for review)

ACKNOWLEDGMENTS

Thanks to all my colleagues at I-Mind Consulting.

CONTACT INFORMATION

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

Linda Palm Simonsson
Magle Stora Kyrkogata 7A
SE-223 50 Lund, Sweden
Work Phone: +46 46 10 25 05
Email: linda.simonsson@i-mind.se

Web: www.i-mind.se