PHUSE EU CONNECT 2021

Paper SI08

How to be Traceable in ADaM

Angelo Tinazzi, Cytel Inc., Geneva, Switzerland

ABSTRACT

Traceability is crucial in all steps of clinical data handling, from data collection to final analysis. For analysis, good traceability starts with a clever design of ADaM datasets, where both data and metadata traceability should be guaranteed.

The CDISC ADaM Team should release in 2021 a document where additional examples of traceability will be provided; furthermore, the importance of traceability in ADaM in data submission, has been recently highlighted by the FDA and PMDA at CDISC EU 2021 ("CDER's experience with the ADaM traceability assessment and common data quality issues", Jesse Anderson).

Retaining --SEQ variable from SDTM source dataset is a starting point, together with other techniques recommended in the CDISC ADaM IG i.e., keeping original variables (and records) used in the derivation.

The purpose of this presentation is to re-enforce the importance of traceability in ADaM through some real examples; moreover, some programming techniques will be illustrated on how to electronically check that data traceability is satisfied, from both SDTM to ADaM and within ADaM i.e., variables copied from ADSL onto other ADaM datasets.

(RE)INTRODUCING TRACEABILITY

The importance of traceability has been already stressed in a number of presentations [1][2][3][4].

For those that are not yet familiar with the concept of traceability, I will use the CDISC definition and an excerpt from the FDA Study Data Technical Conformance Guidance:

> CDISC: "*Traceability is the property that enables the understanding of the data's lineage and/or the relationship between an element and its predecessor(s). Traceability is built by clearly establishing the path between an element and its immediate predecessor*"
>
> FDA: "*Establishing traceability is one of the most problematic issues associated with any data conversion. If the reviewer is unable to trace study data from the data collection of subjects participating in a study to the analysis of the overall study data, then the regulatory review of a submission may be compromised*"

The last part of the sentence from the FDA, looks like, if not yet a threat, at least a warning, but actually more than once FDA members have sent clear message in public speech about how important is for them traceability, and one of them said it is even more important than the standards themselves.

After all, the two definitions give a rationale of why traceability is needed, what we should be able to trace back, either and how this can or should be done.

HOW TO ACHIEVE TRACEABILITY

Technically speaking, in ADaM, Traceability can be achieved in two ways: through *data-point* and through *metadata*.

Datapoint Traceability

Datapoint Traceability by for example simply copying the --SEQ variable from SDTM source dataset, like the example provided in figure 1, where a record in the ADAE is linked to its predecessor record in the SDTM AE dataset by keeping in the ADAE dataset the AESEQ variable from the SDTM AE dataset.

Metadata Traceability

Together with data-point traceability, or whenever data traceability is complex to implement, we can achieve traceability through proper use of metadata, that essentially means to have a good define-xml, for example an algorithm used for deriving a variable properly described in the appropriate section of the define-xml. In the define-xml example provided in figure 2, the value-level metadata portion in the "Source/Derivation/Comment" column, provide the detailed description in plain English of the derivation applied, together with reference to datasets, variables and selection criteria applied; with that, the reviewer will be able to understand what we have done and he or she will be able to eventually reproduce what we did by applying the same criteria.

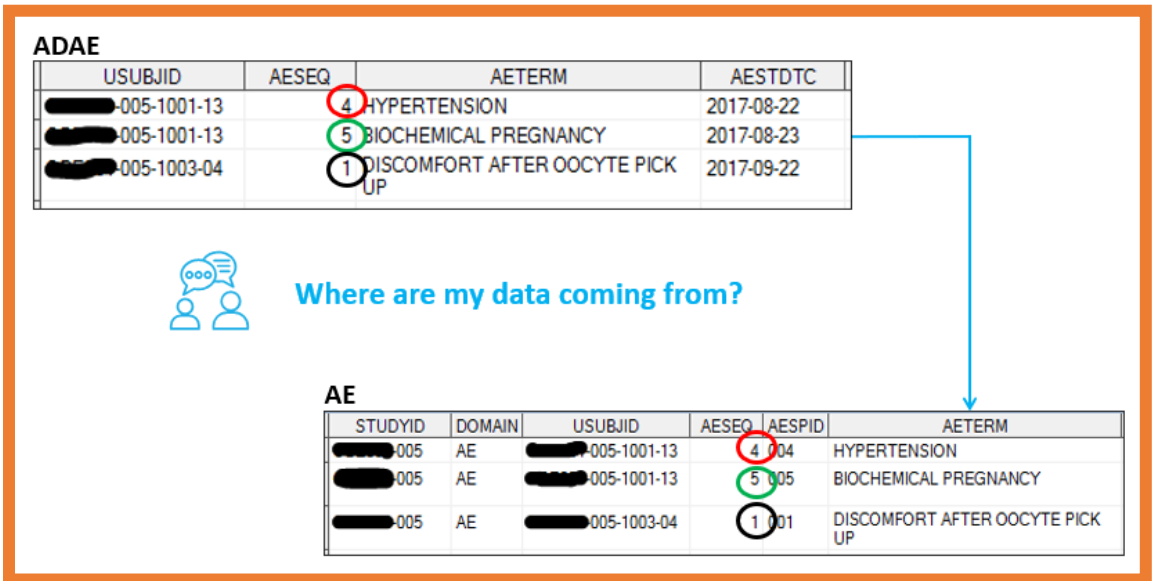


Figure 1: Datapoint Traceability

Parameter Value List - ADTTE [AVAL]

Variable	Where	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
AVAL	PARAMCD = "TTIBG360" (Time to Increased Blood Glucose Levels (Hours) 1-360)	integer	8		Derived: Time to the earliest significantly increased blood glucose level in hours was derived as the time from administration of study treatment (ADSL.TR01SDT) to the first blood glucose level (AVAL when ADBGL.PARAMCD = 'BGLAVGHR') that is ≥ 2 SDs above the 72-hour baseline average blood glucose level (AVAL when ADBGL.PARAMCD = 'BGLAVGDL' and AVISITN = 2 and BASETYPE = '72HRS'). AVAL is the relative hour (ATPTN) of first increased blood glucose level till 360 Hours. Patients without a ≥ 2 SD increase during the time period was censored at the end of the period.

Figure 2: Metadata Traceability

The ADaM IG provides several examples on what other methods you can use to maintain data traceability in your ADaM datasets package. If an ADaM dataset is well designed by applying the applicable ADaM rules and best practices, then by just opening an ADaM dataset a reviewer should, to a certain extent, understand what kind of transformation have been made to which data. This can be achieved through other more sophisticated data-point traceability techniques. For example:

- when you need to **re-derive** in ADaM a variable collected or derived in SDTM, for example the age, you should **create a new variable and keep in the ADaM dataset the original variables from which the ADAM variable was derived**

- similarly, when you need to **derive a variable using variable(s)** from SDTM, you should **keep the original SDTM variables used in the derivation**

- when you need to **impute a missing or a partial value**, for example a date, you should **create a new variable in ADaM and keep the original SDTM variable. For date/time variable you should also make use of imputation flag variables** (e.g., ADTF), so that it is clear which part of the date was imputed

- when you need to apply time windowing methods, for example visits, you should create an analysis visit variable for example and the **keep "supporting" ADaM windowing variables and original SDTM timepoint variables**, for example the SDTM variables VISITNUM and VISIT

- when you need to **impute a missing timepoint from other timepoints**, for example a missing visit, you should derive a new record, and **you should keep the original timepoints records even if not analysed. You should also in this case make use of standards ADaM flags to identify records/visits used in the analysis versus those not used**

- when you deal with **complex derivations requiring several steps, it is recommended to make use of Intermediate datasets** [5][6].

Table 1 summarizes the available « tools » we can make use to support traceability. This includes tools where you can provide additional information, either through metedata with define-xml and CRF, but also with additional documentations specifically developped to support cases where both data-point and metadata traceability are not feasible or difficult to implement, such as the Analysis Data Reviewer Guide (ADRG).

	DATA POINT	METADATA SUPPORTIVE DOCUMENTS
ADaM	• Copy/ retain SDTM variables • Copy/ retain SDTM records • --SEQ from SDTM • SRCDOM/SRCVAR/SRCSEQ • ADTF • ASEQ • DTYPE • ANLxxFL • Occurrence Flags in OCCDS • Intermediate ADaM Datasets	• define.xml • ADRG • SAP
Analysis Results	« Not applicable »	• define.xml (ARM extension) • ADRG • SAP

Table 1: How to be "Technically" Traceable in ADaM (Summary)

ADAM TRACEABILITY EXAMPLES

The ADaM IG contains a number of examples of best practices to make sure our ADaM datasets are traceable enough. To provide more use cases, the CDISC ADaM Team should release in Q4-2021 a document where additional examples of traceability will be provided. The document has already gone through the public review process and its content presented in a couple of public events.

The document that was out for review earlier this year and, among other general recommendations such as what is the optimal number of analysis datasets, how to be analysis ready, etc., it contains 12 instances of fully explained metadata samples together with outputs and ADaM datasets examples such as:

- General ADSL Traceability (see example in figure 3)
- Traceability with Parameters from Multiple Input Datasets, for example when the derivation of a study endpoint requires variables and records from multiple input SDTM or ADaM datasets
- Traceability When Multiple Datasets are Merged
- Traceability When Multiple Input Datasets are Stacked to Create OCCDS
- Traceability When Using a Look-Up Table for example when in addition to SDTM and ADaM datasets our derivation makes use of non-subject data to facilitate the derivation/classification based on criteria requiring lengthy if-then-else
- Traceability When Adding a Row to a BDS Dataset (see example in figure 4)
- Traceability When Multiple Analysis Variables are needed on the Same Row, for example for statistical analysis methods featuring multiple independent/dependent variables
- Using an Intermediate Dataset for BDS Traceability / Using an Intermediate Dataset for ADSL Traceability, when dealing with complex and articulated derivations [5] [6]

GENERAL ADSL TRACEABILITY

Let's assume we want to derive in ADSL, because we need it in our demographics table, the age in categories as per SAP. I would need to get from the SDTM dataset the date of birth and randomization date and the age. It could be also that the SAP has a definition of age that does not match the age either entered or derived in SDTM, in that case we would need to create a new variable age in ADSL; however, for traceability purpose and for the ADaM fundamental principles we will not change the value of the original SDTM AGE variable, instead we will create a new AAGE variable from which then we will derive our age in categories; if instead we are happy the way the age has been either collected or derived in SDTM, the variable age in categories will make use of the original AGE variable from SDTM (figure 3).

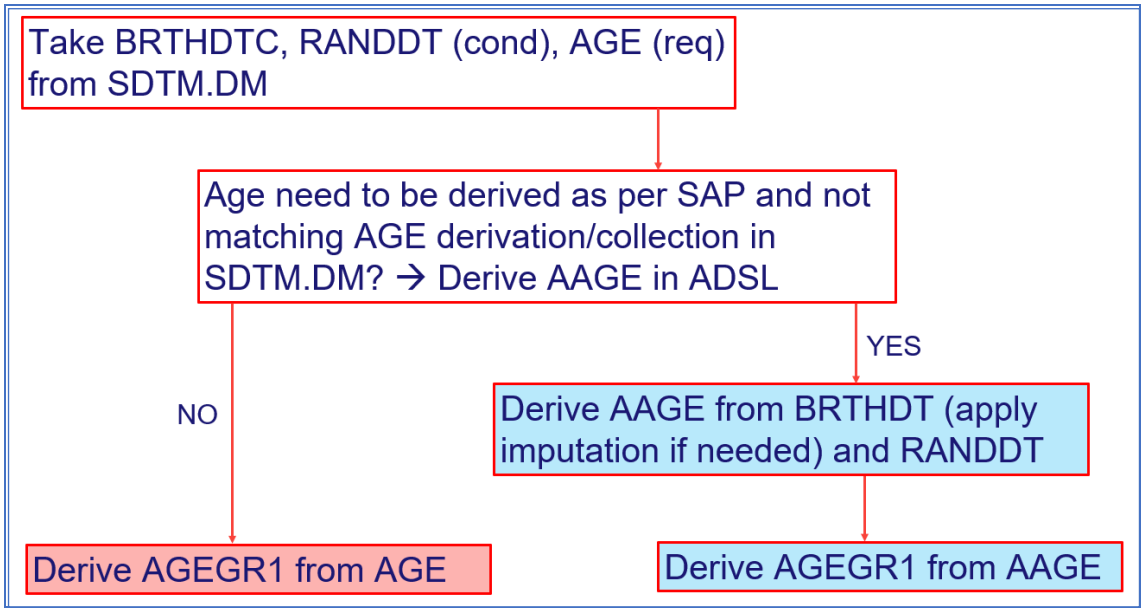


Figure 3: General ADSL Traceability: Derive Age in Categories

CREATING OBSERVATIONS IN BDS

In the example in figure 4, we need to derive a study primary efficacy endpoint defined as the *weekly mean pain intensity score*. In doing that, we already applied some windowing criteria where we have defined the analysis visit, AVIST, based on the relative study day.

In deriving the weekly average, we will need to exclude observations where the subject was still in the wash-out phase, where essentially the subject was still making use of other drugs to cure its pain; the observations occurred in the wash-out period are kept in the ADaM dataset, as well as any other original record from SDTM, but flagged with an analysis flag (so in this case ANL01FL will be Null). Instead, to represent the weekly daily pain average score, a new record, a new analysis parameter, is derived instead of a new variable (this is to preserve the standard BDS structure). Of note here, for the "ADaM purist", we could have avoided the use of DTYPE and simply describe in the define-xml the algorithm used.

Since we want to be sure what we did it is fully traceable and we want to be transparent with the reviewer with the methods used to transform our data, in the define-xml we also provide the derivation method used to derive our new parameter (figure 5).

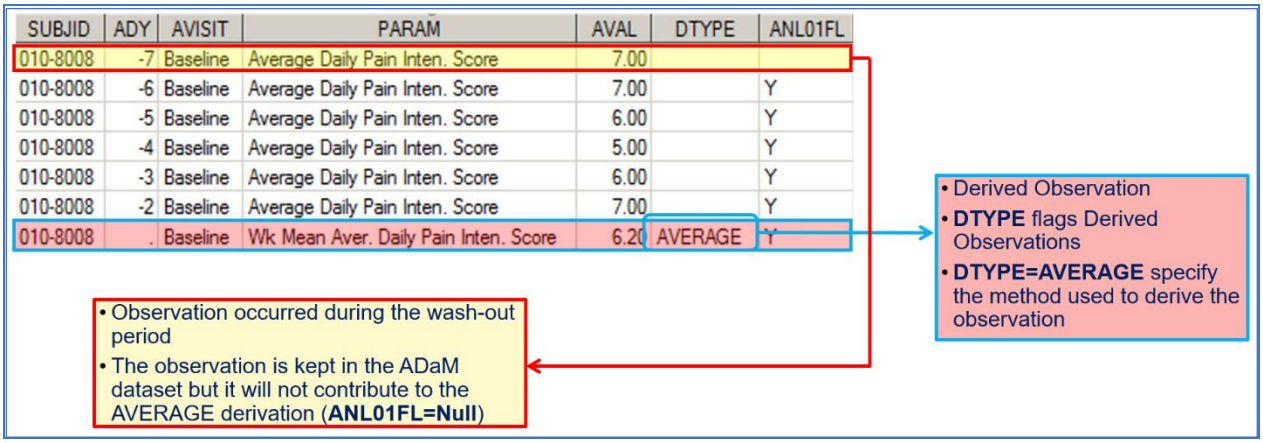


Figure 4: BDS Traceability: Creating new Observations in BDS

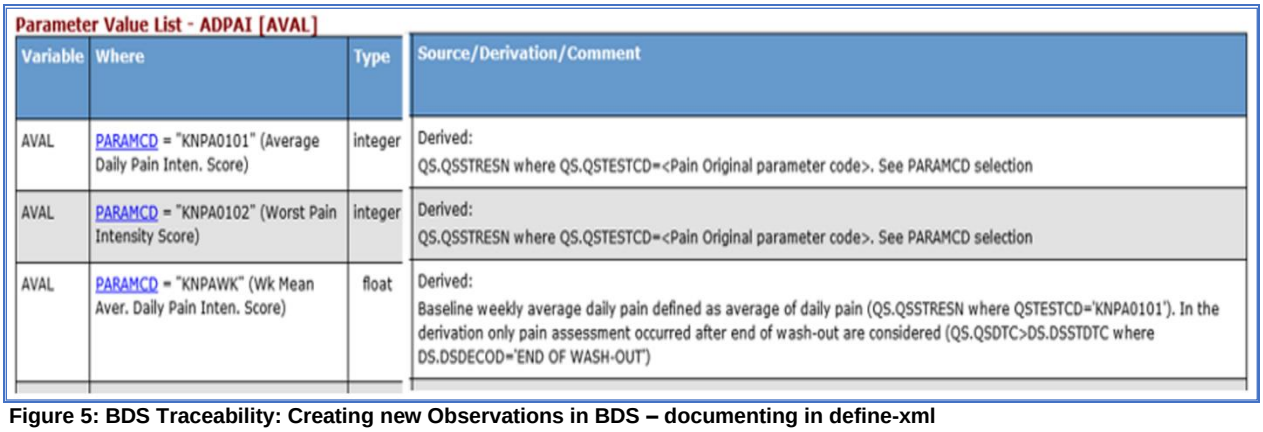


Figure 5: BDS Traceability: Creating new Observations in BDS – documenting in define-xml

IN YOUR REVIEWER SHOES

When creating data submission packages, for example for ADaM, we need to try to be in the shoes, or brain, of the reviewer, and anticipate as much as possible its thinking and therefore the steps he or she will follow in assessing our packages by trying to reproduce what we did.

Analysis Result Metadata (ARM) integrated in the ADaM define-xml, even if not mandatory by all Health Authorities, can be a good start for our good traceability objective, where at least key analysis results are described by specifying the object of the analysis, the analysis variables used and why not the SAS code snippet (figure 6), in this case, with

the exact details of the analysis method used, so that the reviewer will be able to use the same or find alternative approaches to “challenge” what we did.

Furthermore, because we really want to be sure what we did is clear enough, we could provide more or similar details in the analysis data reviewer guide, highlighting the fact that, as in the example in figure 7, our ADPP dataset contains the primary and secondary endpoints. Moreover, we also want the reviewer to be aware that one derived variable, the geometric mean, was calculated in one analysis program rather in the source ADaM dataset.

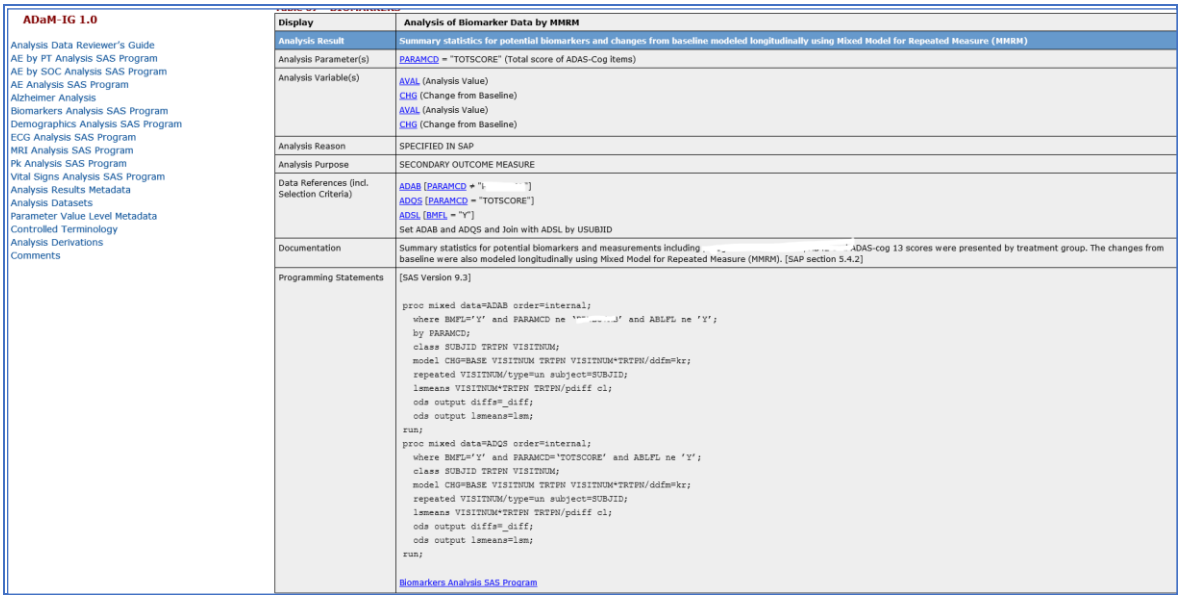
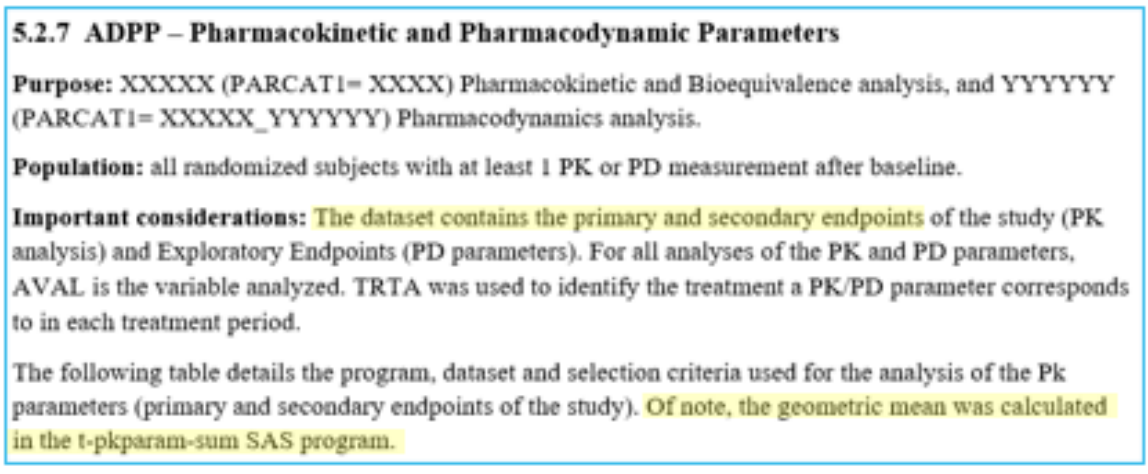


Figure 6: Traceability through the Analysis Results Metadata



5.2.7 ADPP – Pharmacokinetic and Pharmacodynamic Parameters

Purpose: XXXXX (PARCAT1= XXXX) Pharmacokinetic and Bioequivalence analysis, and YYYYYY (PARCAT1= XXXXX_YYYYYY) Pharmacodynamics analysis.

Population: all randomized subjects with at least 1 PK or PD measurement after baseline.

Important considerations: The dataset contains the primary and secondary endpoints of the study (PK analysis) and Exploratory Endpoints (PD parameters). For all analyses of the PK and PD parameters, AVAL is the variable analyzed. TRTA was used to identify the treatment a PK/PD parameter corresponds to in each treatment period.

The following table details the program, dataset and selection criteria used for the analysis of the Pk parameters (primary and secondary endpoints of the study). Of note, the geometric mean was calculated in the t-pkparam-sum SAS program.

Figure 7: Traceability through the Analysis Data Reviewer Guide – Where is the primary endpoint?

In alternative to ARM, a good approach would be, as per the specific sponsor analysis data reviewer guidance template in figure 8, to describe the key study endpoints analysis by providing and describing the source analysis programs and which ADaM datasets have been used and eventually the applied selection criteria.

HEALTHCARE AUTHORITIES CARE ABOUT TRACEABILITY

Often Health Authorities during public event, such as the CDISC Interchange, provided feedback on the quality of data packages received. One very famous example, back in 2012 and updated afterward in few instances, are the "FDA CDER/CBERs Top 7 CDISC Standards Issues"; there the focus was more on the SDTM packages.

More recently at the 2021 CDISC EU Interchange [7], FDA has also shared some very detailed assessment of ADaM traceability in the submitted ADaM packages they have received so far from various sponsors and again, in their conclusions, they have stressed the fact that some of their findings in the traceability assessment they made could call for questions and might result in an information request letter to be sent to the applicant sponsor. In summary another warning from the agency!

Endpoints	Analysis Dataset and Selection Criteria	SAS program name
Primary endpoint: Cmax and AUC_{∞}	ADPP and PKFL='Y' and and ANL01FL="Y" and the following PARAMN/PARAM: ▪ 1 / AUC Infinity (h*pg/mL) ▪ 7 / Cmax (pg/mL)	t-pk-var1-pkpop-sas.txt t-pk-var1-safpop-sas.txt
Secondary endpoint: Cmax, AUC_{∞} and other Pk parameters	ADPP and PKFL='Y' and parcat1='XXXXXX' and ANL01FL="Y" and the following PARAMN/PARAM: ▪ 2 / AUC0-336 (h*pg/mL) ▪ 3 / AUC0-504 (h*pg/mL)	t-pkparam-sum-sas.txt t-pk-var2-pkpop-sas.txt t-pk-var2-safpop-sas.txt

Figure 8: Traceability through the Analysis Data Reviewer Guide – Which programs contain analysis of primary endpoint?

VERIFY TRACEABILITY PROGRAMMATICALLY

We often stress, during the CDISC ADaM training, the importance of the assessment of the ADaM datasets conformance to the standard. I would not discuss it here, but it is important to say that a lot of the conformance assessment in ADaM depends on the subjective assessment, that means we need a solid process to internally validate and review, hopefully independently, the ADaM datasets, the define-xml and the analysis data reviewer guide.

As for the "objective" assessment, with all its standards, CDISC provides the so-called "Conformance Rules", through which we can create tools to validate the conformance of our data packages.

With regards to the ADaM conformance rules pertaining to the verification of the ADaM traceability, CDISC has defined 13 checks that essentially make some basic verifications between some ADaM datasets and some key SDTM datasets, such as DM, EX, AE, ADSL and ADAE (figure 9).

This is of course not enough, and more checks should be put in place to verify traceability and, I would say, consistency for between SDTM and ADaM packages. For example, a check that is often not made, and that is not available in the current standard CDISC validation tools, it is the consistency between SDTM and ADaM, by making sure for example that what is reported in the ADaM define-xml is correct. A typical example is when the origin of an ADaM variable is "Predecessor", meaning that, in this case the variable was copied as it is from the source SDTM dataset. We therefore need to make sure the variable attributes were not changed, as well its content, but most important is that the variable exists in the predecessor SDTM dataset, so that also the predecessor variable attribute was not misspelled in the ADaM define-xml [8]. Of course, probably, a good metadata repository, if well implemented, should guarantee such a traceability between SDTM and ADaM.

Things like derivation methods, would be instead more complicated to be verified programmatically, so here a manual independent review is still needed.

Another example of traceability we could easily check programmatically it is when variables from ADSL are copied onto other ADaM datasets; again, the same we discussed before for SDTM predecessor should be guaranteed between

ADaM datasets and the example in figure 10 could be easily implemented in SAS to make sure ADSL variables when copied to other ADaM datasets are not altered.

- AD0053 (Error) - The combination of STUDYID and USUBJID value does not exist in the SDTM **DM** domain
- AD0061 (Error) - SDTM.**EX** is present but neither ADSL TRTSDT nor TRTSDTM are present
- AD0061A (Error) - SDTM.**EX** is present but neither ADSL TRTEDT nor TRTEDTM are present
- AD0204 (Error) - For the same USUBJID, the ADSL.AGE != **DM**.AGE
- AD0205 (Error) - For the same USUBJID, the ADSL.AGEU != **DM**.AGEU
- AD0206 (Error) - For the same USUBJID, the ADSL.SEX != **DM**.SEX
- AD0207 (Error) - For the same USUBJID, the ADSL.RACE != **DM**.RACE
- AD0208 (Error) - For the same USUBJID, the ADSL.SUBJID != **DM**.SUBJID
- AD0209 (Error) - For the same USUBJID, the ADSL.SITEID != **DM**.SITEID
- AD0210 (Error) - For the same USUBJID, the ADSL.ARM != **DM**.ARM
- AD0253 (Error) - Record key from SDTM **AE** is not traceable to ADaM ADAE (not enough ADAE recs)
- AD0258 (Error) - Record key from ADaM ADAE is not traceable to SDTM.**AE** (extra ADAE recs)
- AD0367 (Error) - For the same USUBJID, the ADSL.ACTARM != **DM**.ACTARM

Figure 9: ADaM Conformance Rules verifying SDTM to ADaM Traceability

```
%*--- Check Variable in ADSL same in other data;
%MACRO ADSL_CHK(INDS=,SOURCE=,SOURCEWH=);

  title1 "Checking obs in ADSL vs ADaMs";
  title2 "&INDS";

  /*Variable in ADSL also in &INDS*/
  proc sql noprint;
    select distinct a.name into :KEEP separated by ' '
    from metadata a, metadata b
    where a.name=b.name and a.memname='ADSL' and b.memname=upcase("&INDS");
  quit;

  /*Sort ADSL*/
  proc sort data=idad.ADSL(keep=&KEEP) out=ADSL_compare;
    by USUBJID;
  run;

  /*Nodupkey and sort &INDS*/
  proc sort data=idad.&INDS.(keep=&KEEP) out=&INDS._compare nodupkey;
    by &KEEP;
  run;
  proc sort data=&INDS._compare nodupkey;
    by USUBJID;
  run;

  /*Compare*/
  title3 "Compare Content";
  proc compare data=ADSL_compare(drop=ADSDTM) compare=&INDS._compare(drop=ADSDTM) listall MAXPRINT=30000;
    id STUDYID USUBJID;
  run;

%MEND ADSL_CHK;
```

Figure 10: Traceability checking for variables copied from ADSL onto other ADaM datasets

CONCLUSIONS

Good ADaM traceability could avoid misunderstanding and questions from the reviewer.
When traceability is successfully implemented, it is in fact possible for the reviewer to identify things like information that exists in the submitted SDTM datasets, information that is derived or imputed within the ADaM dataset, the method used to create derived or imputed data, and Information, records, visits, etc. used for analyses, in contrast to information that is not used for analyses but yet included in the ADaM dataset to support traceability or future analysis.
Traceability could also of course facilitate automation of our processes.

REFERENCES

1. « Lost » in Traceability, from SDTM to ADaM …. finally, Analysis Results Metadata, A. Tinazzi; CDISC EU Interchange, 2016
2. Lost in Traceability- From SDTM to ADaM, A. Tinazzi, Cytel Blog; https://www.cytel.com/blog/lost-in-traceability
3. Traceability: Some Thoughts and Examples for ADaM Needs, S. Minjoe, Q. Zhou and R. Watson; PharmaSUG, 2018
4. More Traceability: Clarity in ADaM Metadata and Beyond, W. Zhong, R. Watson, D. Ewing and J. Zhang; PharmaSUG, 2019
5. ADaM Intermediate Dataset: how to improve your analysis traceability, A. Tinazzi, T. Curto and A. Aggarwal; PHUSE US Connect, 2018
6. Leveraging Intermediate Data Sets to Achieve ADaM Traceability, Y. Zhuo; PharmaSUG, 2019
7. CDER'S Experience with the ADaM Traceability Assessment and Common Data Quality Issues, J. Anderson; CDISC EU Interchange, 2021
8. Please Take Care of your Metadata, A. Tinazzi; PHUSE EU Connect, 2020

CONTACT INFORMATION

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

Angelo Tinazzi
Cytel Inc.
Route de Prè-Bois 20
1215 Geneva – Switzerland
Email: angelo.tinazzi@cytel.com
Web: www.cytel.com

Subscribe to Cytel Blog to receive update on my **Good Data Submission Doctor** Blog Series
https://www.cytel.com/blog/topic/cdisc