PharmaSUG 2017 - Paper SS03

Use of Traceability Chains in Study Data and Metadata for Regulatory Electronic Submission

Tianshu Li, Celldex Therapeutics, Hampton, NJ

ABSTRACT

Traceability is one of the fundamental requirements for electronic submission. It helps the FDA or other regulatory agencies to understand the data's lineage and the relationships among the process of the data collection, SDTM and ADaM data generation, Metadata, and analyses results. The establishment of a clear and unambiguous traceability chain will show the transparency of the electronic submission (e-sub) package and build confidence in the quality of the analyses results and statistical conclusions. Ultimately, it will help expedite the review and approval process.

Based on oncology data as an illustration, this paper describes the type, elements and relationships of good traceability chains, and some important considerations in the process.

INTRODUCTION

Typically, there are two types of traceability chains in the e-sub package; the Metadata traceability chain and the data point traceability chain:

- Metadata traceability chain: Documents enable the reviewers to understand the results of analyses, derivation rules and algorithms, statistical models, source data and variables used, etc. Examples include Analyses Results Metadata and Define.xml of SDTM data and ADaM data.
- Data point traceability chain: specific variables help the reviewers to go directly to the data and/or variable predecessors, or the data collection instruments. Examples include the variables SRCDOM, SRCSEQ and SRCVAR in ADaM data; and the variables _SEQ, VISIT, and VISITNUM in SDTM data.

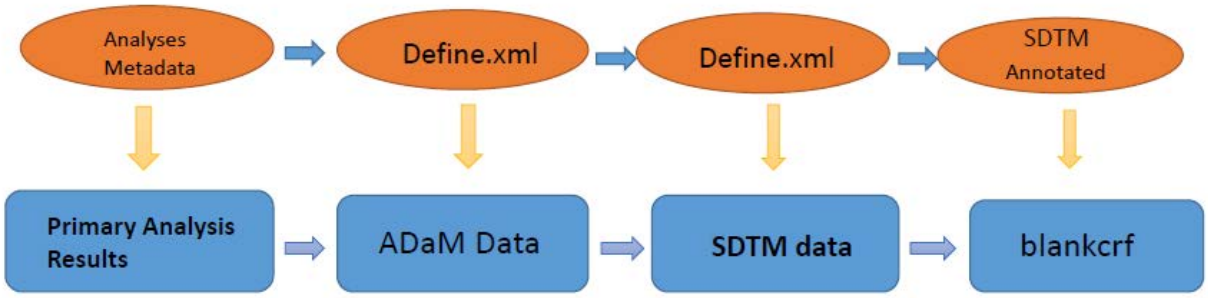

Figure 1: Traceability Chains for Electronic Submission

Figure 1 shows the two traceability chains: the top is the metadata traceability chain and the bottom is the data point traceability chain. The two chains are in parallel in the same process direction, and each chain can function independently. On the other hand, each metadata traceability component describes a particular data point element (yellow arrow) and supplements the traceability functions for the data

point element. Together, the two chains establish the traceability function which is one of the essential principles in the CDISC standards.

The concept of the two traceability chains is illustrated below, based on a typical oncology study efficacy endpoint of ORR (Objective Response Rate) according to the Response Assessment in Neuro-Oncology Criteria (RANO). There are many different therapeutic areas and situations, however, the concept of the traceability chains present in this paper can still be applied in different therapeutic areas.

The examples presented in this paper follow the published CDISC standards. For example, the Analyses Results Metadata follow CDISC publication *Analysis Results Metadata v1.0 for Define-XML v2*; while define.xml examples follow CDISC *define_xml_2_0_releasepackage20140424*.

EXAMPLE OF A CSR PRIMARY EFFICACY TABLE

This table is a portion of a CSR table with two treatment arms. The statistics in the two columns are counts and percentages. The last row is the p-value, which is the primary efficacy interest.

Table 14.2.4.1

Best Overall Response – Response Evaluable Population -Primary Endpoint

	Treatment A (N=xx)	Treatment B (N=xx)
Best Overall Response n (%)		
CR	Xx (xx.x)	Xx (xx.x)
PR	Xx (xx.x)	Xx (xx.x)
SD	Xx (xx.x)	Xx (xx.x)
PD	Xx (xx.x)	Xx (xx.x)
NE	Xx (xx.x)	Xx (xx.x)
Objective Response Rate (CR+PR) n (%)	Xx (xx.x)	Xx (xx.x)
p-value from CMH test	x.xxxx	

This p-value is so important; it may well decide the fate of the whole application. There is no doubt the reviewer(s) would run the statistical model and get a p-value by themselves. However, before that, the reviewer(s) probably will need to fully understand the source data and the data chains to the original data collection point. This is not an easy task. The reviewer(s) may go directly to the metadata traceability component, the define.xml; or the data point traceability element, the ADaM data itself. However, since this is the primary endpoint, the best place to reference is the Analyses Results Metadata.

Submission of the Analyses Results Metadata is recommended by the CDISC for the primary/secondary efficacy endpoints. For other endpoints, the submission of the Analyses Results Metadata may/may not be required, depending on the importance of the statistical conclusions. For example, it may be useful to submit the Analyses Results Metadata for important safety endpoints.

Table 1 shows an example of the Analyses Results Metadata. It provides references for Table 14.2.4.1, and lists important information on how the p-value for the primary endpoint was generated. For traceability propose, the data source is the ADEF, the data selection variables are PARSCA1 and PARAMCD, namely, PARSCA1='RANO' (The Response Assessment in Neuro-Oncology Criteria) and

PARAMCD='ORR' (Objective Response), and the analyses variable is the AVALC. By following the Analyses Results Metadata, a statistician or other reviewer(s) could easily go to the ADEF data, follow the information in the Analysis Results Metadata, and replicate the statistical process which generated the p-value.

Table 1: Analyses Results Metadata

Table 14.2.4.1

Display	Table 14.2.4.1 Best Overall Response – Response Evaluable Population -Primary Endpoint
Analysis Results	p-value for Objective Response Rate (CR+PR) – primary efficacy endpoint
Analysis Parameter(s)	PARAMCD='ORR', objective response
Analysis Variable(s)	AVALC, analysis value of objective response
Analysis Reason	Specified in SAP
Analysis Purpose	Primary efficacy analysis
Data References (incl. Selection Criteria)	ADEF [PARSCA1='RANO' and PARAMCD='ORR']
Documentation	1. ORR is defined as the proportion of patients with evaluable disease who achieve an objective response (a confirmed CR or PR according to RANO criteria assessed by IRC). Best response of CR or PR must be confirmed at least four weeks apart. 2. p-value is from the CMH test, stratified by the randomization factors (updated after randomization) for comparison of CR/PR rate between two treatment groups. 3. SAS® FREQ procedure was used to calculate the p-value 4. Refer to SAP Section 10.2 for derivation rules

Programming Statements	ods output CMH=cmh;
	proc freq data=source;
	tables factor1*factor2*factor3*trt01p* avalc/ cmh scores=modridit;
	run;

DATA POINT TRACEABILITY ELEMENT: ADAM DATA ADEF

By the CDISC standard, whenever possible, the ADaM data should clearly establish the path between an element and its immediate predecessor(s), so its value(s) can be traced back to the data source. The variables SRCDOM, SRCVAR and SRCSEQ variables are the most common traceability variables in ADaM data for BDS (Basic Data Structure). The three variables define the source data name, variable name and the sequence number in the source data, respectively.

Tables 2 lists the example of the three variables, since the ADEF is the BDS. For paramcd='BOR' or 'ORR', the data source was the SDTM data RS, the variable was RSSEQ, and the value of RSSEQ was 5.

Table 2: Snapshot of ADEF Data

SRCDOM	SRCVAR	SRCSEQ	TRT01P	PARAM	PARAMCD	PARSCA1	AVALC	ADY
RS	RSSEQ	5	Treatment A	Best Overall Response	BOR	RANO	CR	156
RS	RSSEQ	5	Treatment A	Objective Response Status	ORR	RANO	Y	156

METADATA TRACEABILITY COMPONENT: DEFINE.XML FOR ADEF

The define.xml is the Metadata to describe the data ADEF. It provides information on how each and every variable was derived, its predecessor(s), variable characteristics, etc.

Table 3: An example of define.xml for ADaM data

ADEF-Primary Efficacy Analyses [Location: adef.xpt]

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	text	12		Predecessor: ADSL.STUDYID
SRCDOM	Source Data	text	12		Assigned
SRCVAR	Source Variable	text	12		Assigned
SRCSEQ	Source Sequence Number	integer	8		Predecessor: RS.RSSEQ
ADT	Analysis Date	integer	8		Derived: SAS date from RS.RSDTC
PARAM	Parameter	text	100	PARAM_ADEF	Assigned:
PARAMCD	Parameter Code	text	8	PARAMCD_ADEF	Assigned:
PARAMN	Parameter (N)	integer	8	PARAMN_ADEF	Assigned: Assign a numeric code for each value of PARAMCD (see codelist PARAMN_ADEF)
AVALC	Analysis Value	text	20		Derivations are described per parameter in the parameter value level metadata
TRT01P	Planned Treatment	text	20	<ARM>	Predecessor: ADSL.TRT01P

Usually, for traceability propose, two levels of naming conversions are used to name a predecessor data and variable: a data domain name, followed by a variable name. For example, ADSL.STUDYID.

Table 3 includes several traceability variables: the predecessor of STUDYID variable in ADEF was the STUDYID variable in ADaM data ADSL; and the predecessor of SRCSEQ variable in ADEF was the RSSEQ variable in SDTM data RS. The values of the traceability variables SRCDOM and SRCVAR are assigned and provided the information on the predecessor data and variables for a particular PARAMCD.

DATA POINT TRACEABILITY ELEMENT: THE RS DATA

Table 4 shows the snapshot of the RS data. There are two traceability variables in the example: the RSSEQ, which was referred in the ADEF variables SRCDOM, SRCVAR and SRCSEQ. And the VISIT variable, which was the linkage between the RS data and the SDTM annotated CRF.

Table 4: Snapshot of RS Data

VISIT	RSSEQ	RSDTC	RSTEST	RSTESTCD	RSORRESC
WEEK 11	5	2015-02-15	Overall Response	OVRLRESP	CR

Table 5: An example of define.xml for SDTM

Tumor Response (RS) [Location:RS.XPT]

Variable	Label	Key	Type	Length	Controlled Terms or Format	Origin	Derivation/Comment
STUDYID	Study Identifier	1	text	7		Protocol	
DOMAIN	Domain Abbreviation		text	2	["RS" = "Tumor Response Data"] <Domain Abbreviation (RS)>	Assigned	
USUBJID	Unique Subject Identifier	2	text	14		Derived	Concatenation of STUDYID and SUBJID
RSSEQ	Sequence Number		integer	2		Derived	Sequential number identifying records within each USUBJID in the domain.
RSTESTCD	Tumor Assessment Short Name	5	text	7	Tumor Assessment Code	Assigned	
RSTEST	Tumor Assessment Name		text	22	Tumor Assessment Name	eDT	
RSORRES	Result or Finding in Original Units		text	8		eDT	
VISIT	Visit Name		text	7		eDT	

Table 5 is a snapshot of the define.xml for SDTM RS data. The STUDYID variable was referred in the dfine.xml for ADEF as a predeceesor, The traceability variables RSSEQ was derived in the RS data, and referred in the define.xml for ADEF variables SRCDOM, SRCVAR and SRCSEQ; while the VISIT was the data collected in the SDTM annotate CRF.

SDTM ANNOTATED CRF (BLANKCRF.PDF)

For legacy or, maybe other reasons, the CRF name is referred to as the blankcrf.pdf. But it's not blank at all. Because a real blankcrf.pfd does not contain the information on how the CRF field were mapped to a particular SDTM data and the variables, so it's very difficult, if not impossible, to trace back from the SDTM data to a CRF source data. In practical, only the SDTM annotated CRF can serve this propose, and should be used in the electronic submission to perform the traceability function.

Table 6: An example of SDTM annotated CRF

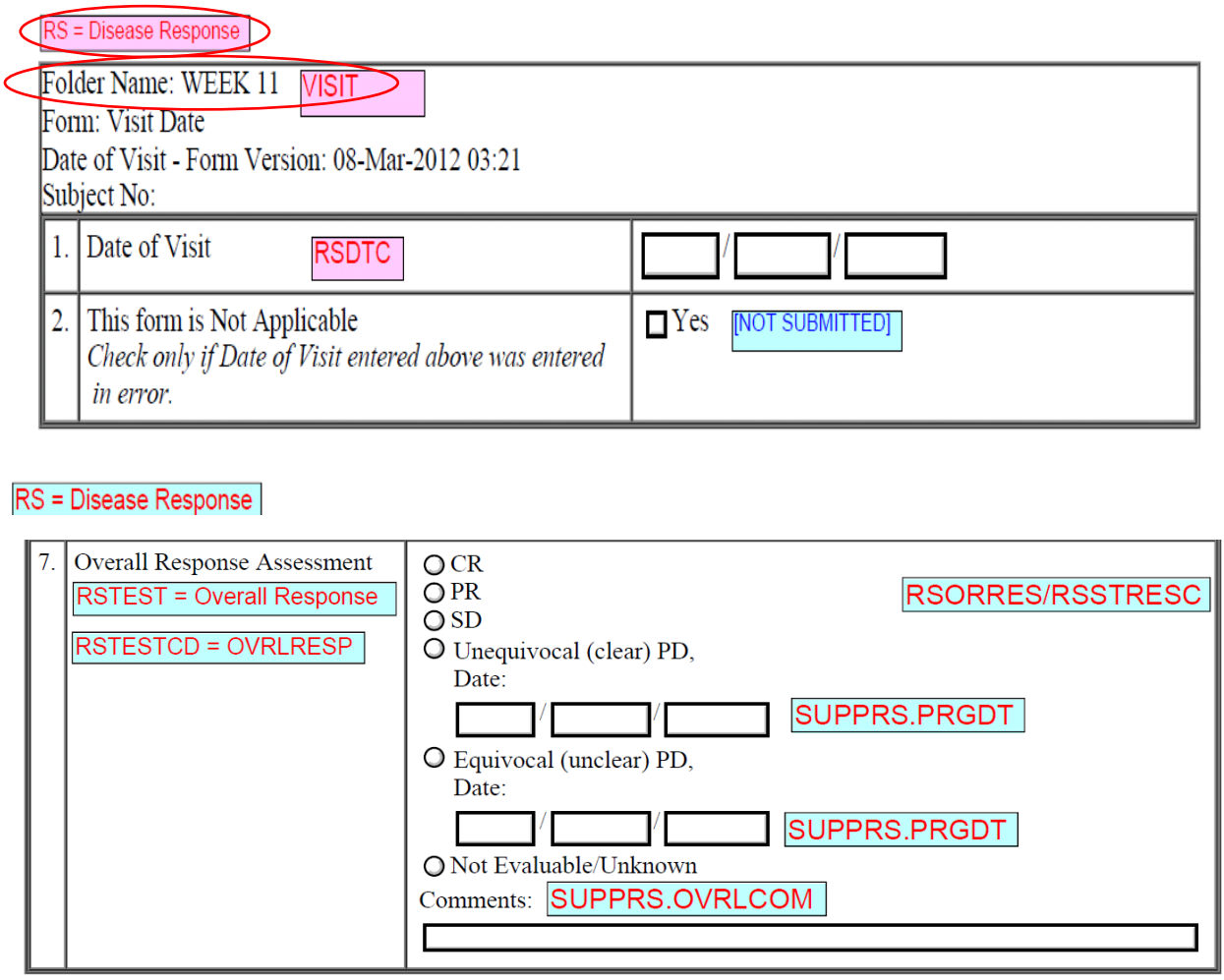

Table 6 provides an example of an SDTM annotated CRF for the RS data. RS is the data name, and value of the Folder Name is 'WEEK 11' and it was mapped to the VISIT variable in RS. Other variables in the RS data can be traced back to the SDTM annotated CRF also. For example, the RSDTC, RSTEST, RSTESTCD and RSORRESC can be located in the CRF.

SPECIAL CONSIDERATIONS

1. **Why both the Data Point Traceability chain and the Metadata Traceability chain are needed?**

It might appear redundant to have two traceability chains repeating the same kind of information in parallel. However, there are good reasons behind it: the metadata traceability chain supplements and enhances the data point traceability chain.

By the CDISC standards, the data traceability is one of the fundamental data building blocks. As it's critical that anyone who reviews the data can fully understand the data well, it's warranted to define them both. Secondly, although the data point traceability chain offers a self-explaining nature of the traceability for the data, sometimes the traceability cannot be established by the data point traceability chain alone due to data structure restrictions.

For example, in general no traceability is provided for the ADSL data structure in the data point traceability chain. Another example is that there may be situations where complex variable derivations are required. Under such conditions, many source variables maybe used to derive a variable, and it is not practical to define the traceability in the data itself, or very difficult to describe them in the data. The only option is to define them in the metadata traceability chain. For example, in the define.xml, which is much more flexible, the traceability can be explained by free text if needed.

Under all situations, a clear and unambiguous traceability must be defined, through either the data point or metadata traceability chain, or both, to meet the CDISC and the submission standards.

2. **Intermediate Traceability data**

To make the traceability concept easy to understand, this paper uses a simple case as the example. In real submission packages, there may be complex situations where the Data Point Traceability is very difficult, if not impossible, to define. For example, time-to-event analyses are very common for oncology studies. It's not easy to define the traceability in the ADTTE data. In this situation, intermediate traceability data may be needed to form a bridge collecting the ADTTE and the source data, and establish the data point traceability chain. (A detailed discussion of the intermediate traceability data is beyond the scope of this paper).

The submission of the intermediate data is not required at this point. However, it's a good practice and could be very helpful for the reviewer(s) to understand the data if they are included in the e-sub package.

For metadata traceability chain, the parameter value level definition in define.xml (version 2.0) can nicely define the traceability for ADTTE or other BDS (Basic Data Structure) data.

3. **Non-standard source data**

The non-standard source data refers to any collected data which is not defined in the CRF. It can be the non-standard data mapped to the SDTM data, or a data source for ADaM data. There are no standard

ways defined in the Data Point traceability arena. So, the only way is to go through the Metadata Traceability chain path. Usually, a link to the definition file (define.xml or define.pdf) of such data in the SDTM or ADaM define.xml file is necessary and acceptable.

4. **Study Data Reviewer's Guide or Analysis Data Reviewer's Guide.**

When a traceability function does not belong to a standard traceability chain, or if it's an important factor in determining the outcome of a major endpoint, the data traceability could also be explained in the two Reviewer's Guides.

CONCLUSION

The two traceability chains and the two Reviewer's Guides establish the complete traceability functions for an e-sub package. This paper uses a very simple example to demonstrate the basic concept of the traceability. However, in real world, situations can become very complex. One can always follow the CDISC standards, including the tools presented in this paper, along with good and knowledgeable judgement, to achieve the goal of clear and unambiguous data traceability.

ACKNOWLEDGMENTS

I would like to thank Jennifer Green and Yi He who reviewed this paper and gave me valuable comments

CONTACT INFORMATION

Should you have any comments and questions, please contact:

Tianshu Li
Celldex Therapeutics
53 Frontage Road, Suite 220
Hampton, NJ 08827

tianshuli@celldex.com