PharmaSUG 2019 - Paper DS-185

Leveraging Intermediate Data Sets to Achieve ADaM Traceability

Yun (Julie) Zhuo, PRA Health Sciences

ABSTRACT

Traceability, a fundamental principle of ADaM, provides transparency and increases confidence for the FDA reviewers. Building traceability could be a daunting task for a complex ADaM data set especially when it involves multiple data sources and multi-step derivations. In this paper, we illustrate the benefits of using an intermediate data set to achieve traceability using example SDTM and ADaM data sets from a Phase III oncology study. Through the intermediate data set, along with its metadata, it is possible to trace the final analysis value to a record in the intermediate data set, and then from there to the source domains.

INTRODUCTION

WHAT IS TRACEABILITY

According to the Analysis Data Model Implementation Guide (ADaMIG) v1.1, traceability is defined as:

"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."

Simply put, ADaM traceability is a clear path that we build to ensure the results we create can be traced back to its source SDTM data sets.

WHY IS TRACEABILITY IMPORTANT

Traceability facilitates understanding, ensures transparency, and increases confidence in the results. Traceability is one of the fundamental principles of ADaM. Unfortunately, broken traceability is not uncommon between ADaM and its predecessors, especially where complex derivations are involved. According to the FDA Study Data Technical Conformance Guide (SDTCG):

"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."

HOW TO BUILD TRACEABILITY

Before we dive into complex examples, we are going to quickly review the simple traceability methods and concepts provided in the ADaMIG. We are going to refer to the following concepts throughout this paper.

- Traceability has two levels: metadata traceability and data point traceability.

  Metadata traceability is established by describing via metadata the algorithm used or steps taken to derive or populate an analysis value from its immediate predecessor. Metadata is required for ADaM compliance.

  Data point traceability is established by providing clear links in the data (usually via the use of SRCDOM, SRCVAR, and SRCSEQ variables) to the specific data values used as input. Data point traceability should be implemented if practically feasible.

- Traceability establishes across-dataset relationships as well as within-dataset relationships.

  An example of the across-dataset traceability will be the SRCDOM, SRCVAR, and SRCSEQ variables which provides the link between output data set and its source domain.

  An example of the within-dataset traceability will be the metadata for the analysis flag variables, which enables the users and reviewers to understand how the records were derived.

ILLUSTRATIVE STUDY EXAMPLE

THE ANALYSIS NEEDS

Prior and subsequent cancer therapies are important information collected in oncology trials. They are objects of interest in many parts of the analysis plan in our illustrative study example.

Progression Free Survival

In oncology trials, a common efficacy endpoint is progression free survival (PFS). In our illustrative example, a hypothetical Statistical Analysis Plan (SAP) states:

"PFS is defined as the time from randomization to disease progression or death, whichever occurs first. If the subject has a subsequent cancer therapy prior to the event, then the subject will be censored at the last tumor assessment date prior to the subsequent cancer therapy. Subjects alive and not meeting the criteria for progression will be censored at the last evaluable tumor assessment. If the subject has no evaluable assessment, then the subject will be censored at the date of randomization. If the subject has any subsequent cancer therapy, then the subject will be censored at the last evaluable assessment date prior to the first subsequent cancer therapy."

The corresponding ADaM data set for PFS analysis will be created in the time-to-event data structure (ADTTE), as shown below.

USUBJID	PARAM	PARAMCD	AVAL	STARTDT	ADT	CNSR	EVNTDESC	CNSDTDSC
S01-001	Progression Free Survival (days)	PFS	439	01JAN2018	15MAR2019	0	DEATH	
S01-002	Progression Free Survival (days)	PFS	244	01JAN2018	01SEP2018	0	PROGRESSION	
S01-003	Progression Free Survival (days)	PFS	486	01JAN2018	01MAY2019	1	COMPLETED STUDY	LAST TUMOR ASSESSMENT SHOWING NO PROGRESSION
S01-004	Progression Free Survival (days)	PFS	121	01FEB2018	01JUN2018	2	EARLY DISCONTINUATION	LAST TUMOR ASSESSMENT SHOWING NO PROGRESSION
S01-005	Progression Free Survival (days)	PFS	90	01FEB2018	01MAY2018	3	SUBSEQUENT CANCER THERAPY	LAST TUMOR ASSESSMENT PRIOR TO SUBSEQUENT CANCER THERAPY
S01-006	Progression Free Survival (days)	PFS	1	01FEB2018	01FEB2018	4	NO BASELINE ASSESSMENT	RANDOMIZATION

Table 1. Example Time-to-Event Data Set (ADTTE) for PFS Analysis

The first row in Table 1 indicates subject S01-001 has a death event, and the second row indicates subject S01-002 has a progression event. It implies that both subjects do not have any subsequent cancer therapy before their specific event occurrence date. A comparison of event date with subsequent cancer therapy date, if there is any, is necessary.

The row with purple highlight has an event description of subsequent cancer therapy. Subject S01-005 was censored at the last tumor assessment date prior to subsequent cancer therapy. To determine the analysis date for censoring (ADT), we need to compare first subsequent cancer therapy date with tumor

assessment dates. ADT is defined as last tumor assessment date prior to first subsequent cancer therapy, on the condition that there is no event (disease progression or death) prior to any subsequent cancer therapy.

Descriptive Summary

In our hypothetical analysis plan, we are also interested in a descriptive tabular display of all prior and subsequent cancer therapies, grouped by categories.

Table xxx. Prior and Subsequent Cancer Therapies

	ARM A (N=XX)	ARM B (N=XX)	Overall (N=XX)
Any Prior Cancer Therapy	xx (xx%)	xx (xx%)	xx (xx%)
Systemic Cancer Therapy	xx (xx%)	xx (xx%)	xx (xx%)
Radiotherapy	xx (xx%)	xx (xx%)	xx (xx%)
Surgery	xx (xx%)	xx (xx%)	xx (xx%)
Any Subsequent Cancer Therapy	xx (xx%)	xx (xx%)	xx (xx%)
Systemic Cancer Therapy	xx (xx%)	xx (xx%)	xx (xx%)
Radiotherapy	xx (xx%)	xx (xx%)	xx (xx%)
Surgery	xx (xx%)	xx (xx%)	xx (xx%)

Table 2. Example Shell Table for Prior and Subsequent Cancer Therapies

Protocol Deviation Analysis

Our hypothetical SAP also states:

“All important deviations related to study inclusion or exclusion criteria, conduct of the trial, patient management or patient assessment should be described. Protocol deviations should be appropriately summarized and grouped into different categories. “

Receiving cancer therapy while subject is still on study drug treatment is one of the protocol deviations commonly analyzed in oncology trials. The corresponding example ADaM dataset for the analysis is shown below:

USUBJID	DVCAT	DVTERM	DVSTDT
S01-010	Prohibited Therapy	Systemic cancer therapy (RCHOP) while on study drug.	01MAR2018
S01-011	Eligibility Criteria	No baseline tumor assessment.	02JAN2018
S01-012	Product Administration Error	Received less than target dose.	01SEP2018

Table 3. Example ADDV for Protocol Deviations Analysis

It will be the focus of this paper to discuss the generation of ADaM compliant analysis data sets for all the cancer therapy analysis needs in the study, with the goals of improving efficiency, simplifying process, and more importantly, laying down a clear, transparent traceability path for the users (reviewers, validation programmers, statisticians). We will demonstrate that the intermediate data set technique brings significant advantages.

THE COMPLEXITY EXPLAINED

The basis of the intermediate data set technique is the fact that very complex derivation algorithm blurs the traceability path. In this section, we will explain the complexities involved in the cancer therapy analysis.

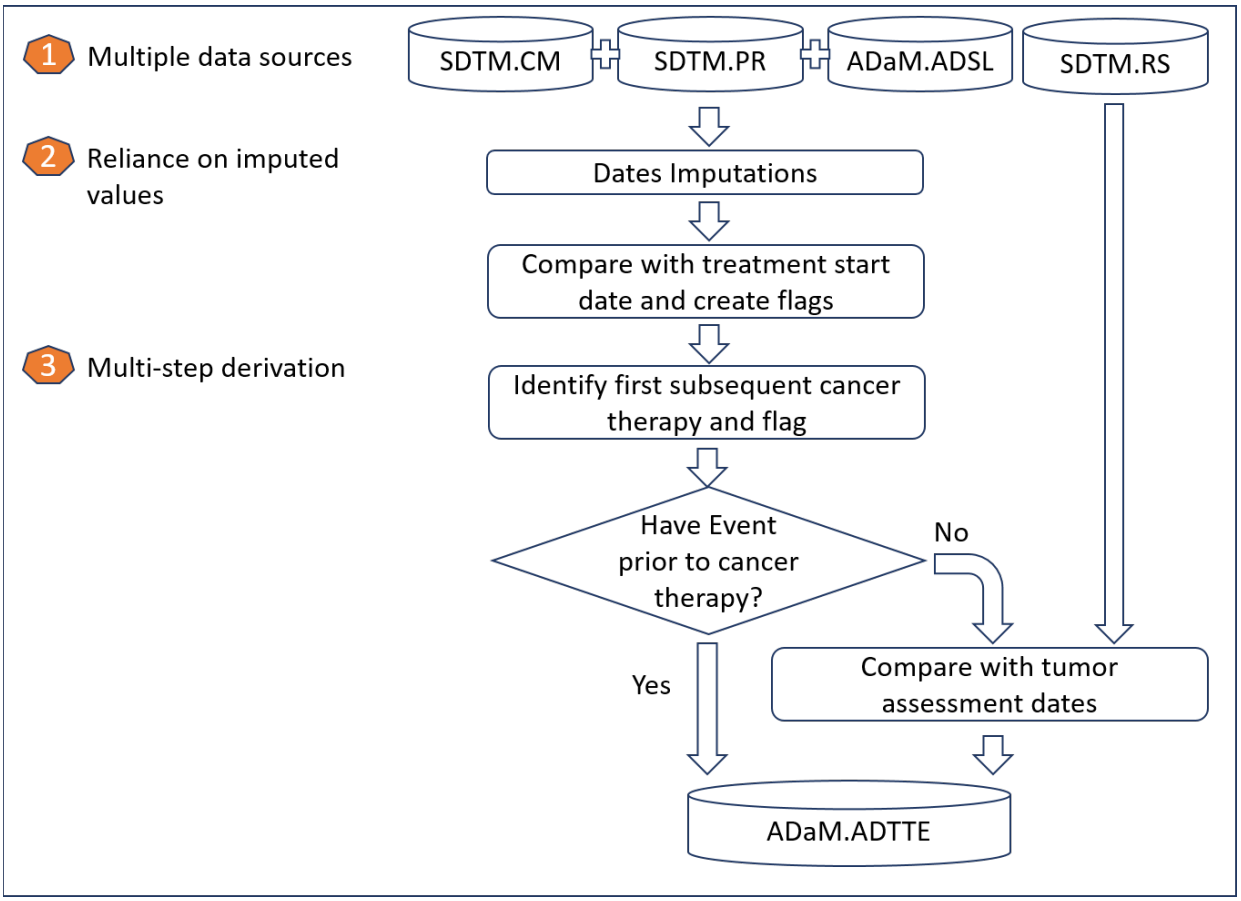

Figure 1. **Data Derivation Flow for Cancer Therapy Analysis**

Figure 1 above reveals three levels of complexities involved in the algorithm.

1. There are multiple data sources. Cancer therapies data are stored in the SDTM CM and PR domains. From the SDTM CM domain, we obtain data records for all systemic cancer therapies. From the SDTM PR domain, we obtain data records for radiotherapies and surgeries.
2. The derivation relies on imputed date values. Cancer therapy dates are often missing or partially missing. We apply imputation rules on missing or partially missing records, and we rely on imputed values for later date comparisons.
3. It takes multiple steps to derive. In addition to the step of date imputation, we compare cancer therapy dates with treatment start date, event (progression or death) start date, and tumor assessment dates. We also need to identify specific records for analysis.

As a result, creating traceability directly from the source SDTM to final analysis data set will be a daunting task for the programmers, and a confusing result for the users.

THE INTERMEDIATE DATA SET

To address the analysis needs and ensure traceability, we are going to create our first intermediate data set which we name it ADCT (Analysis Data Set for Cancer Therapies). Since it includes a well-structured hierarchy of categories and terminology, it fits nicely into the ADaM Occurrence Data Structure (ODS).

We firstly assembles all prior and subsequent cancer therapies from two source domains: SDTM.CM and SDTM.PR. Example source data sets are shown below. Only relevant content are included here.

Domain	USUBJID	CMSEQ	CMTRT	CMDECOD	CMCAT	CMSTDTC
CM	S01-002	1	RCHOP	RCHOP	SYSTEMIC CANCER THERAPY	2018-12-01
CM	S01-005	1	Interferon 2B	INTERFERON	SYSTEMIC CANCER THERAPY	2015
CM	S01-005	2	Interferon ALFA	INTERFERON	SYSTEMIC CANCER THERAPY	2018-06

Table 4. Example Data: Cancer Therapies from the SDTM CM Domain.

Domain	USUBJID	PRSEQ	PRTRT	PRCAT	PRSTDTC
PR	S01-002	1	RADIOTHERAPY	RADIOTHERAPY	2010-06
PR	S01-002	2	HYSTERECTOMY	SURGERY	2019-01
PR	S01-005	1	RADIOTHERAPY	RADIOTHERAPY	2018-07-15

Table 5. Example Data: Cancer Therapies from the SDTM PR Domain.

Secondly, we obtain a number of relevant variables from ADSL in order to perform date imputation and date comparison. In our example data below, we included treatment start date (ADSL.TRTSDT), death date (ADSL.DTHDT), and first progression date (ADSL.FPDDT). The intermediate data set ADCT with only relevant content is shown in Table 6 below.

USUBJID	ASEQ	TRTSDT	DTHDT	FPDDT	ACAT	ATRT	ASTDT	ASTDTF	SRCDOM	SRCSEQ	ANL01FL	ANL02FL	ANL03FL	ANL04FL
S01-002	1	01JAN2018		01SEP2018	RADIOTHERAPY	RADIOTHERAPY	01JUN2010	D	PR	1	Y			
S01-002	2	01JAN2018		01SEP2018	SYSTEMIC CANCER THERAPY	RCHOP	01DEC2018		CM	1		Y	Y	
S01-002	3	01JAN2018		01SEP2018	SURGERY	HYSTERECTOMY	01JAN2019	D	PR	2		Y		
S01-005	1	01FEB2018		10OCT2018	SYSTEMIC CANCER THERAPY	Interferon 2B	01JAN2015	M	CM	1	Y			
S01-005	2	01FEB2018		10OCT2018	SYSTEMIC CANCER THERAPY	Interferon ALFA	01JUN2018	D	CM	2		Y	Y	Y
S01-005	3	01FEB2018		10OCT2018	RADIOTHERAPY	RADIOTHERAPY	15JUL2018		PR	1		Y	Y	

Table 6. Example Data: Cancer Therapies Intermediate Data Set

Example dataset-level metadata (Table 7) and variable-level metadata (Table 8) are shown below. Only relevant content are included.

Data Set	Data Set Description	Data Structure	Class of Data Set
ADCT	Cancer Therapies Analysis Data Set	One record per subject per cancer therapy	Occurrence Structure

Table 7. Example Data Set Level Metadata for ADCT

Variable Name	Variable Label	Code List /Controlled Terminology	Source/Derivation
USUBJID	Unique Subject Identifier		Predecessor: ADSL.USUBJID
ASEQ	Analysis Sequence Number		Unique sequential number within each subject
TRTSDT	Treatment Start Date		Predecessor: ADSL.TRTSDT
DTHDT	Death Date		Predecessor: ADSL.DTHDT
FPDDT	First Progression Date		Predecessor: ADSL.FPDDT

ACAT	Analysis Category	"SYSTEMIC CANCER THERAPY" "RADIOTHERAPY" "SURGERY"	For data from SDTM.CM, set to CM.CMCAT For data from SDTM.PR, set to PR.PRCAT
ATRT	Analysis Therapy Name		For data from SDTM.CM, set to CM.CMTRT For data from SDTM.PR, set to PR.PRTRT
ADECOD	Analysis Standardized Therapy Name		For data from SDTM.CM, set to CM.CMDECOD For data from SDTM.PR, set to PR.PRDECOD
ASTDT	Analysis Start Date of Therapy		For data from SDTM.CM, use CM.CMSTDTC For data from SDTM.PR, set to PR.PRSTDTC Convert to numeric date format and apply imputation rules specified in the SAP.
ASTDTF	Analysis Start Date Imputed Flag	"D" = "Day" "M" = "Month" "Y" = "Year"	Assign "D" when Day is imputed, else assign "M" if Month is imputed, else assign "Y" if Year is Imputed.
SRCDOM	Source Domain		For data from SDTM.CM, set to 'CM' For data from SDTM.PR, set to 'PR'
SRCSEQ	Source Sequence Number		For data from SDTM.CM, set to CM.CMSEQ For data from SDTM.PR, set to PR.PRSEQ
ANL01FL	Analysis Flag - 01	"Y" = "Yes"	Prior Cancer Therapy: Flag 'Y' if ASTDT is prior to TRTSDT
ANL02FL	Analysis Flag - 02	"Y" = "Yes"	Subsequent Cancer Therapy: Flag 'Y' if ASTDT is on or after TRTSDT
ANL03FL	Analysis Flag - 03	"Y" = "Yes"	On-treatment Cancer Therapy: Flag 'Y' for subsequent cancer therapies that occurred before treatment discontinuation.
ANL04FL	Analysis Flag - 04	"Y" = "Yes"	First Subsequent Cancer Therapy before Progression or death: If there is no progression or death, flag 'Y' for the first occurrence of subsequent cancer therapy. If there is any progression or death, flag 'Y' for the first occurrence of subsequent cancer therapy prior to disease progression or death.

Table 8. Example Variable Level Metadata for ADCT

The intermediate data set accomplished the following:

1. Combined all cancer therapies from two SDTM domains, which is necessary for selecting interested records and generating flags.
2. Performed imputation on missing and partially missing date values. The imputed portions are highlighted in yellow in Table 6. The imputed date values will be used in multiple analysis downstream for improved efficiency. The ASTDTF variable is added in compliance with the CDISC date imputation rules.
3. Created the SRCDOM and SRCSEQ variables for traceability links with input data sets.
4. Created the ASEQ variable, which is a required variable for intermediate data sets in order to establish a traceability link with the next level of data set/analysis.
5. Four derived analysis flag variables identifying interested records for different types of analysis.

USUAGE AND TRACEABILITY

In this section, we are going to follow the traceability path while demonstrating the three analysis usage of the ADCT data set.

Usage 1: As Final Data Set for Tabular Display

To start with, in our first example, ADCT will serve as the final ADaM data set for a simple descriptive display of prior or subsequent cancer therapies. The traceability path is demonstrated below.

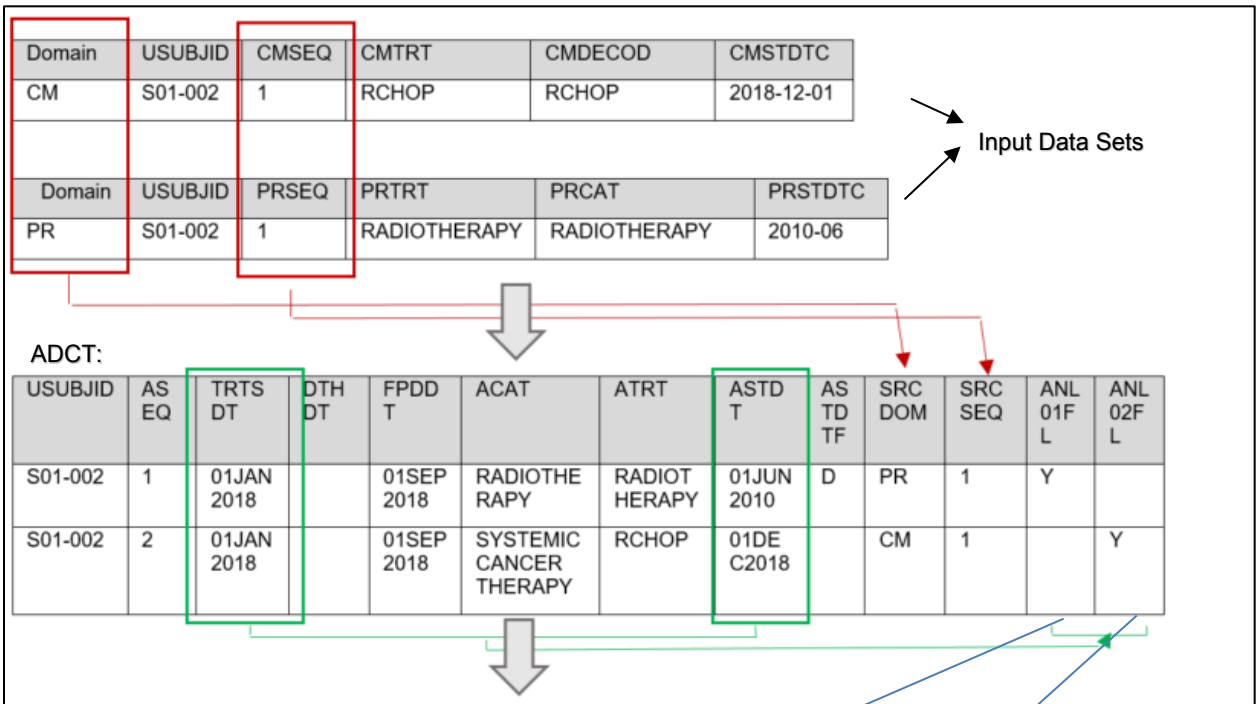

Table xxx. Prior and Subsequent Cancer Therapies

	ARM A (N=XX)	ARM B (N=XX)	Overall (N=XX)
Any Prior Cancer Therapy [ADCT.ANL01FL='Y']	xx (xx%)	xx (xx%)	xx (xx%)
Systemic Cancer Therapy [ADCT.ACAT]	xx (xx%)	xx (xx%)	xx (xx%)
Radiotherapy [ADCT.ACAT]	xx (xx%)	xx (xx%)	xx (xx%)
Surgery [ADCT.ACAT]	xx (xx%)	xx (xx%)	xx (xx%)
Any Subsequent Cancer Therapy [ADCT.ANL02FL='Y']	xx (xx%)	xx (xx%)	xx (xx%)
Systemic Cancer Therapy [ADCT.ACAT]	xx (xx%)	xx (xx%)	xx (xx%)
Radiotherapy [ADCT.ACAT]	xx (xx%)	xx (xx%)	xx (xx%)
Surgery [ADCT.ACAT]	xx (xx%)	xx (xx%)	xx (xx%)

Figure 2. Traceability: Multiple Sources and Derivations in the Occurrence Data Structure

In Figure 2, the data flow in red color demonstrates across-dataset traceability established by domain (CM.DOMAIN and PR.DOMAIN), sequence (CM.CMSEQ and PR.PRSEQ), source domain (SRCDOM), and source sequence (SRCSEQ) variables. The variable SRCDOM provides the name of the input SDTM domain, and the SRCSEQ is carried over from the source domains to link each row in ADCT to its

corresponding row in the source domain. For example, the first record of subject S01-002 in ADCT is carried over from the PR domain where PRSEQ=1.

The data flow in green color demonstrates within-dataset traceability established by analysis flag variables (ANL01FL and ANL02FL) within the ADCT data set. The variable level metadata (Table 9 below) clearly explains the source variables and the derivation logic.

Variable Name	Variable Label	Code List /Controlled Terminology	Source/Derivation
TRTSDT	Treatment Start Date		Predecessor: ADSL.TRTSDT
ASTDT	Analysis Start Date of Therapy		For data from SDTM.CM, use CM.CMSTDTC For data from SDTM.PR, set to PR.PRSTDTC Convert to numeric date format and apply imputation rules specified in the SAP.
ANL01FL	Analysis Flag - 01	"Y" = "Yes"	Prior Cancer Therapy: Flag 'Y' if ASTDT is prior to TRTSDT
ANL02FL	Analysis Flag - 02	"Y" = "Yes"	Subsequent Cancer Therapy: Flag 'Y' if ASTDT is on or after TRTSDT

Table 9. Metadata Traceability for Analysis Flag Variables

In addition, Figure 2 also shows that the analysis flags provide the link between the ADaM data set and the analysis result. The shell table annotations referring to the analysis flags will direct reviewers to the specific records flagged in the ADaM data set.

Usage 2: As Intermediate Data Set for Protocol Deviation Analysis Data Set (ADDV)

The protocol deviations analysis data set (ADDV) contains both collected and derived protocol deviations, compared to the SDTM DV domain which only includes collected deviations, for the purpose of analysis. As discussed previously, receiving subsequent cancer therapy while subject is still on study drug treatment is one of the most common protocol deviations in oncology trials. Similar to ADCT, ADDV is also structured in the ODS format since it identifies the occurrences of sponsor defined deviation events.

Similar to the previous usage example, the across-dataset traceability between SDTM and ADCT is established by domain and sequence variables, and the within-dataset traceability is established by the metadata for the analysis flag variable.

The focus of this usage example is on the traceability path between ADCT and a ODS structured analysis data set ADDV. The traceability is demonstrated in Figure 3.

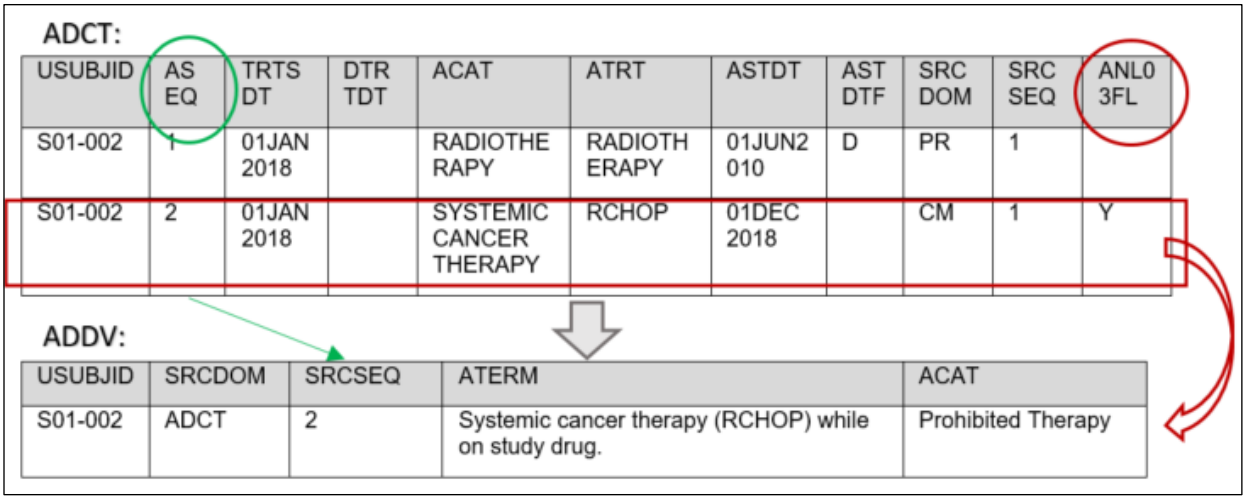

Figure 3. Traceability: From ODS Structured Intermediate Data Set to ODS Structured Final Data Set

Figure 3 shows that, firstly, an analysis flag (ANL03FL) identifies qualified records in ADCT to be carried over to the ADDV data set. Traceability is established by variable level metadata for ANL03FL as described in Table 8.

Secondly, because ADCT is going to be used as the input data set to generate ADDV, it is important that we use a sequence variable named ASEQ in ADCT to provide record level traceability.

Finally, data point traceability variables, including the ASEQ variable in ADCT and the SRCDOM and SRCSEQ variables in ADDV, provides a link between the intermediate data set and the final data set. For example, subject S01-002 has a record of prohibited cancer therapy which is taken from ADCT where the sequence number ASEQ=2.

Usage 3: As Intermediate Data Set for another Intermediate Data Set in Time-to-Event Analysis

As previously demonstrated in Figure 1, the PFS time-to-event analysis has one of the most complex algorithms. With the help of ADCT, we have accomplished the majority of the steps described in Figure 1. Now we need to feed the ADCT intermediate data set into a BDS (Basic Data Structure) structured intermediate data set named ADINTDT, which will then feed into the final ADTTE analysis data set.

Similar to the previous two usage examples, we use the following traceability methods:

1. The across-dataset traceability between SDTM and ADCT is established by data point traceability variables such as domain and sequence variables.
2. The within-dataset traceability is established by the metadata for the analysis flag variable.
3. The across-dataset traceability between ADCT and ADINTDT is established by both data point traceability variables and the analysis flag metadata.

In this usage example, we will focus on the traceability path from ADCT (an ODS structured intermediate data set) to ADTTE (a BDS structured final analysis data set), through another intermediate data set (a BDS structured intermediate data set). Figure 4 demonstrates this process.

Firstly, the analysis flag (ANL04FL) effectively identifies the record of interest, that is, the first occurrence of subsequent cancer therapy. If there is any progression or death event, the occurrence of subsequent cancer therapy has to be prior to any event. Traceability is established by variable level metadata as described in Table 8. Example Variable Level Metadata for ADCT.

Secondly, data point traceability is established by three variables: SRCDOM, SRCVAR, and SRCSEQ. Note that compared to the previous usage example, we added a new variable SRCVAR in order to identify the exact variable that was carried over into the AVALC column of the BDS data structure. For example, the data flow in red color in Figure 4 shows that the second record (ADINTDT.SRCSEQ=ADCT.ASEQ=2) of subject S01-005 has an ASTDT (SRCVAR='ASTDT') value of '01JUN2018', which becomes the corresponding AVALC value in ADINTDT.

Finally, the data flow in green color shows that the last tumor assessment date prior to subsequent cancer therapy will be carried over into ADTTE for subject S01-005. Subject S01-005 is censored at the last tumor assessment date prior to subsequent cancer therapy due to a subsequent cancer therapy event. The derivation of last tumor assessment date in ADINTDT is not discussed in this paper. The link is again established by data point traceability variables (ADTTE.SRCDOM and ADTTE.SRCSEQ).

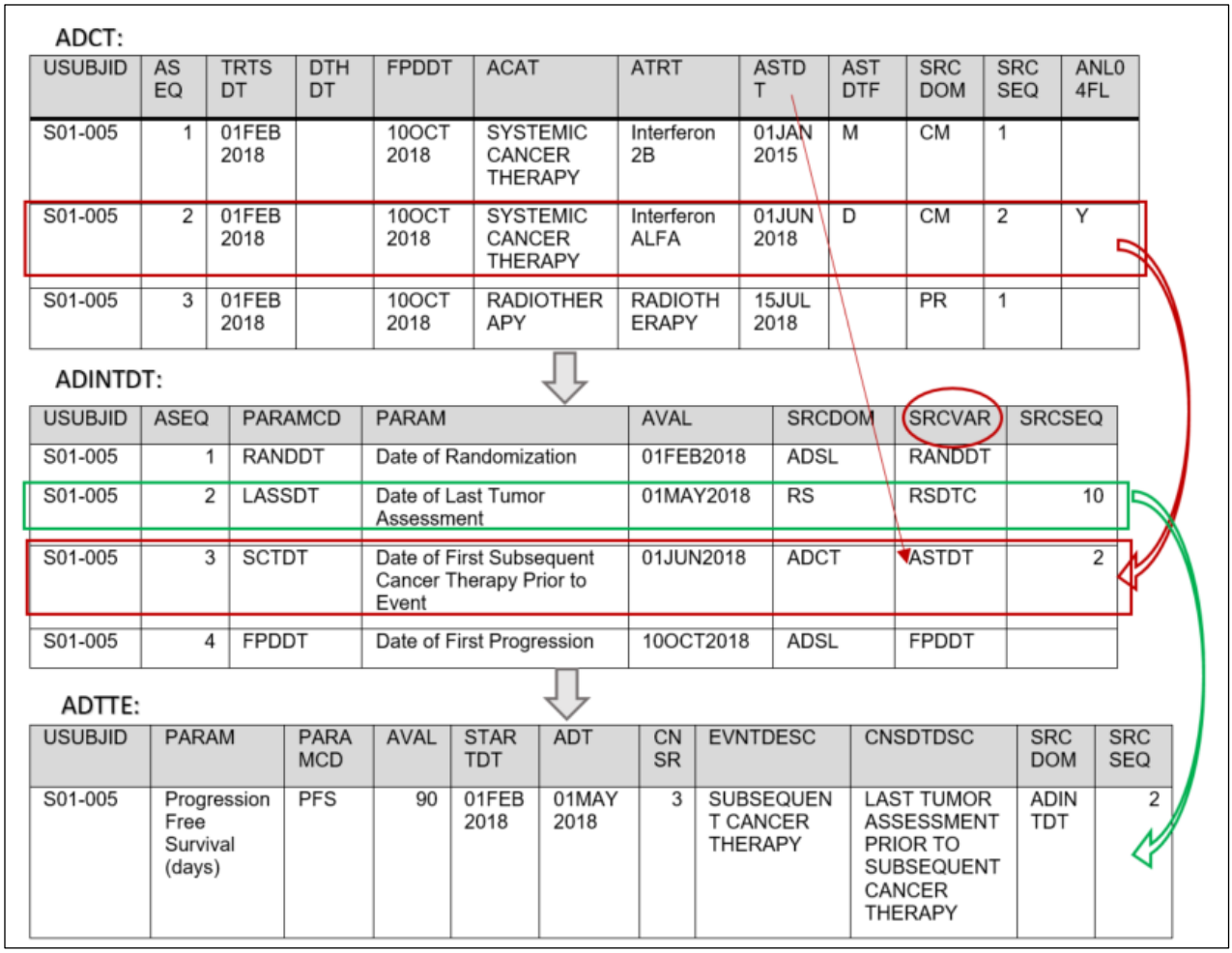

Figure 4. Traceability from ODS Structured ADCT to BDS Structured ADTTE

DISCUSSION

THE CASE FOR INTERMEDIATE DATA SETS

With intermediate data sets, traceability becomes a lot more transparent for complex derivation algorithms.

Firstly, data point traceability becomes clear because it is implemented in a step-by-step manner. Take the time-to-event analysis as example, if we trace the traceability path backward, we follow the SRCDOM and SRCSEQ variables in ADTTE to the corresponding record in ADINTDT from which the analysis date value was carried over. We then follow the triple data point traceability variables (SRCDOM, SRCVAR, and SRCSEQ) in ADINTDT to the record and variable in ADCT from which the analysis value AVALC was carried over. Then, the data point variables in ADCT further lead us to corresponding records in the source SDTM domains.

Secondly, the derivation note in the metadata becomes simpler and easier to understand. For example, without intermediate data sets, the derivation note of the analysis date (ADTTE.ADT) for the subsequent cancer therapy event would have been written like this:

Variable	Where	Source/Derivation
ADT	PARAMCD='PFS'	Subset CM where CMCAT='SYSTEMIC CANCER THERAPY. Subset PR where PRCAT='RADIOTEHRAPY' or 'SURGERY'. Impute CMSTDTC and PRSTDTC (converted to numeric) using imputation rules specified in the SAP. Compare imputed cancer therapy dates with treatment start date (ADSL.TRTSDT), first progression date (ADSL.FPDDT), and death date (ADSL.DTHDT). If there is a subsequent cancer therapy date that is on/after treatment start date and, if there is any event, before FPDDT/DTHDT, then ADT is the latest RS.RSDTC (converted to numeric) where RS.RSTESTCD='OVRLRESP' and RS.RSSTRESC in ('CR' 'PR' 'SD' 'PD') on or prior to the subsequent cancer therapy date.

Table 10. Example Derivation Note in the Metadata without Intermediate Data Set

The lengthy explanation of the derivation logic in Table 10 is confusing and prone to mis-interpretation. By contrast, with intermediate data sets, the derivation logic is de-constructed into separate components. Coupled with data point traceability variables, the derivation logic is easy for reviewers to understand.

Thirdly, the increased transparency benefits not only reviewers but also validation programmers. De-constructed derivation logic is easier for validation programmers to follow, and retained intermediate values make it easier for the programmers to investigate inconsistencies. For example, when the validation output has different result from the production output, validation programmers can easily trace the differences into previous steps and compare the intermediate values in the intermediate data set to detect any error.

In addition, it is of note that, with effective use of analysis flag variables and the horizontal layout, the occurrence data structure in our example is programmatically advantageous for many derivations in the complex algorithm, including comparison of date values and identifying first occurrence of a number of events. A fitting data structure leads to more streamlined programming.

Another important benefit is that, with careful planning, intermediate data sets have the potential to contribute to multiple analysis needs in a study. As a result, we can improve programming efficiency and streamline the programming process. In our example, prior and subsequent cancer therapies are objects of interests to at least three separate analysis. We built in analysis flags so that each transparent subset of the intermediate data set serves each of the analysis purpose.

THE CAUTIONS

It is important that we optimize the number of analysis data sets created for each study. We build intermediate data sets only when there are significant benefits such as facilitating traceability for very complex derivations.

Intermediate data set is not the only tool for traceability needs. For example, ADaMIG section 4.4 includes a less-complex time-to-event example where the analysis values can be traced directly back to multiple source data sets in a BDS structured analysis data set. Therefore, when designing traceability, we want to consider all viable paths before we make a decision on a most suitable method based on the study's specific analysis needs.

CONCLUSION

This paper demonstrates an example where an intermediate data set supporting multiple analysis needs in a study can facilitate traceability so that the final analysis value can be traced to a record in the intermediate data set, and then from there to the source domains. We demonstrate that the data point traceability variables coupled with the metadata are effective tools to communicate the links between data sets. They are even more effective when a complex algorithm is decomposed into separate steps. A wide array of paths are available for traceability, and the selection of a most suitable path depends on the specific needs of the derivations as well as the general picture of the analysis plan.

REFERENCES

CDISC. 2012. “The ADaM Basic Data Structure for Time-to-Event Analyses v1.0” https://www.cdisc.org/system/files/all/standard_category/application/pdf/adam_tte_final_v1.pdf

CDISC. 2016. “ADaM Implementation Guide v1.1” https://www.cdisc.org/standards/foundational/analysis-data-model-adam/adam-implementation-guide-v11

FDA. 2018 “Study Data Technical Conformance Guide” https://www.fda.gov/downloads/forindustry/datastandards/studydatastandards/ucm384744.pdf

Lombardi, E. 2013. “Traceability in the ADaM Standard” *Proceedings of the PharmaSUG 2013 Conference*, Cary, NC: SAS Institute Inc. Available at https://www.pharmasug.org/proceedings/2013/PO/PharmaSUG-2013-PO13.pdf

Minjoe, S. etc. 2018. “Traceability: Some Thoughts and Examples for ADaM Needs.” *Proceedings of the PharmaSUG 2018 Conference*, Cary, NC: SAS Institute Inc. Available at https://www.pharmasug.org/proceedings/2018/DS/PharmaSUG-2018-DS01.pdf

Santopli, T. etc. 2014. “An ADaM Interim Data Set for Time-to-Event Analysis Needs.” *Proceedings of the PharmaSUG 2014 Conference*, Cary, NC: SAS Institute Inc. Available at https://www.lexjansen.com/pharmasug/2014/DS/PharmaSUG-2014-DS09.pdf

Tinazzi, A. etc. 2018. “ADaM Intermediate Data Set: how to improve your analysis traceability.” Proceedings of the PhUSE US Connect 2018, Available at https://www.lexjansen.com/phuse-us/2018/si/SI16.pdf

ACKNOWLEDGMENTS

The author appreciates her former employer Axio Research LLC for the opportunities and guidance in building ADaM analysis data sets.

CONTACT INFORMATION

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

Yun (Julie) Zhuo
PRA Health Sciences
ZhuoYun@prahs.com
Kite Pharma
jzhuo@kitepharma.com