PharmaSUG2011 - Paper TU09

Using the ADaM Basic Data Structure for Survival Analysis

Nancy Brucken, i3 Statprobe, Ann Arbor, MI
Sandra Minjoe, Octagon Research, Wayne, PA
Mario Widel, Roche Molecular Systems, Pleasanton, CA

ABSTRACT

The Clinical Data Interchange Standards Consortium (CDISC) Analysis Data Model (ADaM) team has described a Basic Data Structure (BDS) that can be used for most analyses. This tutorial will walk through how to develop a BDS dataset for survival analysis. Attendees should be familiar with survival analyses and related SAS® procedures, such as PROC LIFETEST.

During the tutorial, attendees will complete by hand some variable-level metadata for an example set of survival analyses. Attendees may find it useful to bring to the tutorial session a copy of the ADaM Implementation Guide (IG), available free from the CDISC website[1] for download.

INTRODUCTION

Although the IG mentions using BDS for survival analysis, most of the examples in that document deal with laboratory data used for change from baseline and shift table analyses. The ADaM team has a subgroup assigned to further describe the use of the BDS structure for survival analysis, and at the time of this writing that draft appendix was under review prior to public posting. However, we can still make use of the rules in the ADaM documents to guide us in creating a BDS dataset that will work for survival analysis.

If, by the time of the conference, the ADaM IG appendix for survival and time-to-event analysis is available for public review, it will be discussed in the tutorial.

ADAM TEAM AND DOCUMENTS

The ADaM team formed in 2000, about a year after the Submission Data Standards (SDS) team. CDISC realized that SDS would not be sufficient for filings because SDS data is designed for data storage and tabulations, not for analysis.

At the time of this writing, the most current ADaM documents are the Analysis Data Model version 2.1 and the ADaM Implementation Guide version 1.0, both published 17Dec2009 and available for free from the CDISC website. ADaM validation checks are also available to member companies. The first ADaM appendix document, ADAE for adverse events, was published in draft form for public review through 18Mar2011. The ADaM appendix document, ADTTE for survival and time-to-event analyses, has not yet been published for public review.

Analysis Process

ADaM is more than just analysis data; it's also analysis results, such as proportions, means, and p-values. The analysis process can be thought of in 2 steps:

1. Create the analysis dataset
2. Create the analysis results

Most of our work is done in step 1, creating the analysis dataset. As described in the ADaM documents, ADaM datasets are to be "analysis ready". This means the dataset doesn't need to be merged with other data, transposed, or have new variables derived before use in a statistical procedure.

All of the work in step 1 allows step 2 to be straightforward. This means that someone with limited programming skills can still use the data. For example, a biostatistician can use an ADaM dataset to create several different ad-hoc analyses, in addition to the ones we created for the study report.

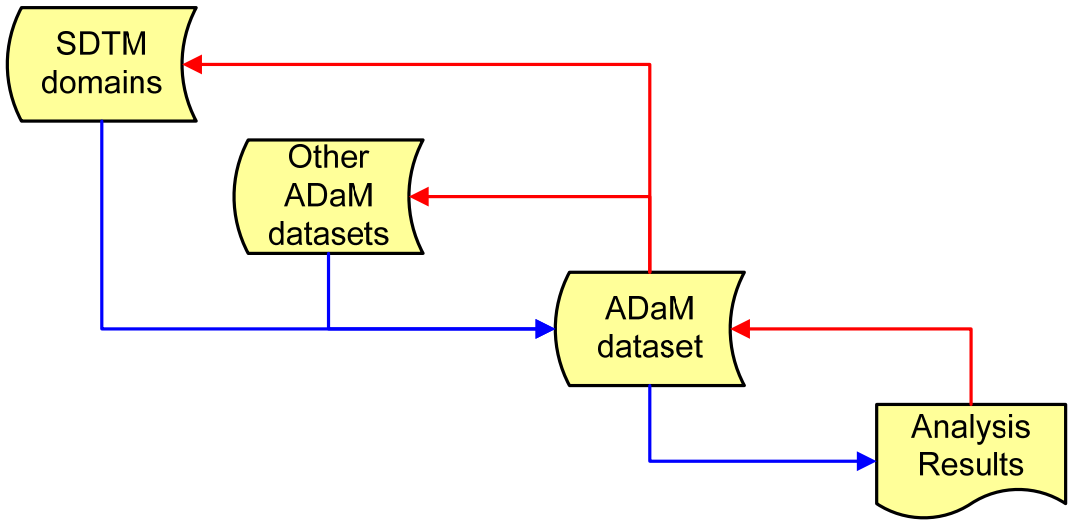
Although we think about these two steps in order of dataset first and then results, in actuality when we're creating specifications we work in the opposite direction. That is,

1. Determine what analysis results are needed
2. Determine what the analysis dataset needs to contain in order to produce the analysis results

3. Determine how to derive the data

In other words, we first need to know what we'll be analyzing before designing our dataset. This allows us not only to create a dataset that is analysis-ready, it also prevents it from being cluttered with a lot of extra data that is not needed for any analyses.

Visually, we can think of these two separate processes, design and analysis, with the following diagram:



The red lines (at the top) show the order in designing our specifications. The blue lines (at the bottom) show the order of running programs to produce our analyses. We first follow the red lines, and then the blue. Thus we actually need to do the following:

1. Determine what analysis results are needed
2. Determine what the analysis dataset needs to contain to produce the analysis results
3. Determine how to derive the data
4. Write/run a program to produce the analysis dataset
5. Write/run the program to produce the analysis results

It isn't uncommon to spend as much or even more time in developing good specifications than on writing and running programs.

SURVIVAL ANALYSIS

Determine What Analysis Results are Needed

As described above, to develop our survival BDS dataset, we first need to understand a little about the analysis needed. Survival analysis is a way of comparing those who had an event happen (such as death) with those who did not. To do this, we need to determine the following:

- Duration of time until the event, such as time from randomization to death. This is simply a difference between the two dates, in whatever unit is necessary for the analysis.
- What to use as a substitute endpoint for those who didn't have the event. This is called censoring, and we'll often use the last date of any information in the study.
- A way to distinguish those who had the event from those who did not. This is just a flag variable.

As described in the ADaM documents, the BDS Structure is one record per subject, per analysis parameter, per time point (as needed). In survival analysis we typically don't need the time point breakdown, since the analysis parameters are generally determined for the entire study.

Determine What the Analysis Dataset Needs to Contain to Produce the Analysis Results

In any BDS structure, the variables PARAM, PARAMCD, PARAMN are used to describe the parameter for analysis. For survival analysis, this might be something like:

PARAM	PARAMCD	PARAMN
Progression-Free Survival	PFS	1
Overall Survival	OS	2
Time to Progression	TTP	3

Variables AVAL (numeric) or AVALC (character) would be the actual value used in the analysis. In survival analysis, we need the numeric version, AVAL, and would set it to the duration. Typically we derive the duration as the difference between the date of the event and a reference start date, such as randomization date. Duration can be in any units, but we might use days, months, etc.

When the event did not happen, e.g., the subject did not die, we censor the event at some date, such as date last known alive, so that every subject in the analysis will have a value in AVAL. Whenever a variable is censored, we want to set a flag to 1. Note: the standard in statistics is to use 1=censored, 0=occurred. SAS may allow us to assign other values, but we should use the standard so that it will be straightforward for the experienced statistician and directly useable within any statistical analysis package.

Determine How to Derive the Data

In order to derive all of these values, we'll need to collect data from our various sources. In many analysis datasets, such as lab change from baseline, we bring in data from ADSL plus one SDTM domain LB. However, with survival analysis we may need data from a variety of datasets, such as:

- DS (disposition) for date of death, date last known alive
- AE (adverse events) or CE (clinical events) for date of a specific event of focus
- LB (laboratory test results) for date of a lab event of focus
- QS (questionnaires) for date of a significant change in status
- Custom domains, such as those used to collect disease status such as progression

We could also pull in data from other analysis datasets, as show in the prior diagram. For simplicity here in this paper, we are using just SDTM domains as input. Here are some selected records for these SDTM domains for one subject:

DOMAIN	DSSEQ	DSSTDTC	DSDECOD
DS	1001	01MAR2010	INFORMED CONSENT OBTAINED
DS	1008	19JUL2010	DEATH

DOMAIN	AESEQ	AESTDTC	AETERM
AE	1012	12APR2010	HEADACHE
AE	1065	19JUL2010	BLURRED VISION

DOMAIN	CESEQ	CESTDTC	CETERM
CE	1021	28APR2010	PAIN IN AREA OF TUMOR
CE	1033	02MAY2010	SWELLING IN AREA OF TUMOR

DOMAIN	LBSEQ	LBSTDTC	LBTESTCD	LBSTRESN	LBTOXGR
LB	1092	06MAY2010	ALB	3.9	1
LB	1093	06MAY2010	ALP	187	4

DOMAIN	QSSEQ	QSSTDTC	QSTESTCD	QSSTRESN
QS	1121	18APR2010	GH	3
QS	1122	18APR2010	GHINDEX	38

DOMAIN	RSSEQ	RSSTDTC	RSTESTCD	RSSTRESC
RS	1013	30JUN2010	TRGRESP	PD
RS	1014	30JUN2010	NTRESP	SD
RS	1015	30JUN2010	OVRESP	PD

For each of these pieces of incoming information, we can create PARAM, PARAMCD and PARAMN values. We won't be analyzing the data in this form, but including them in the dataset allows us to use them to derive our actual analysis parameters and provides traceability. This information can be stored in the ADaM BDS structure, either as part of the actual analysis dataset or as an interim dataset. For simplicity here in this paper, we'll keep it in the analysis dataset.

Thus, as we're pulling together the data from all the SDTM domains, we create the following analysis records for this one subject:

PARAMCD	ADT	AVAL	CENSOR	SRCDOM	SRCVAR	SRCSEQ	DTYPE	EVNTDESC
DEATHDT	19JUL2010	128	0	DS	DSSEQ	1008		DEATH
AEDT	12APR2010	32	0	AE	AESEQ	1012		AE of interest
AEDT	19JUN2010	98	0	AE	AESEQ	1065		AE of interest
CEDT	02MAY2010	50	0	CE	CESEQ	1033		CE of interest
LBDT	06MAY2010	54	0	LB	LBSEQ	1092		Lab test with tox grade of 4 or 5
QSDT	18APR2010	38	0	QS	QSSEQ	1122		QS with GHINDEX < 40
RSDT	30JUN2010	109	0	RS*	RSSEQ	1015		RS overall response result of PD

From this information, we can then derive the variables actually needed for our analysis, by subtracting the reference start date, such as randomization date, from the appropriate date above for each of the following analysis parameters:

PARAMCD	ADT	AVAL	CENSOR	SRCDOM	SRCVAR	SRCSEQ	DTYPE	EVNTDESC*
PFS*	12APR2010	32	1				TTE	First AE, CE, Lab, or Progression of interest
OS*	19JUL2010	128	0				TTE	Death
TTP*	19JUN2010	109	0				TTE	RS overall response of PD

**** Specifications for Progression-free survival, Overall Survival, and Time to Progression may differ. This is not an attempt to teach the statistical theory of survival analysis or suggest definitions for these analysis endpoints, but to show how the dataset can be developed.***

Write/Run a Program to Produce the Analysis Dataset

Once the specifications are complete, we then must write a program to produce the dataset. Programming code is not the focus of this paper.

At time of program development, we may determine it is easier to do this work by first creating interim dataset(s). However, it's worth noting that because we have kept the rows of information for each piece of data that could have contributed to our analysis parameters, we can modify the parameter definitions to use different information without having to change the structure of the dataset.

Write/Run a Program to Produce the Analysis Results

Once we have this information, we can use a procedure such as SAS PROC LIFETEST to do the actual analyses. Again, programming code is not the focus of this paper.

SUMMARY

Although the ADaM team has not yet released the appendix document for survival analysis, using the Analysis Data Model and ADaM Implementation Guide we can create a survival analysis dataset in a BDS structure. We need to know what analyses are needed, so that we can determine the analysis parameters we'll need. And we'll need to know the definitions for both the event and censoring, so that we know the information we'll need to bring in from the SDS domains or other ADaM datasets.

REFERENCES

[1] All ADaM documents can be found on the CDISC website. Public documents are at http://www.cdisc.org/adam.

ACKNOWLEDGEMENTS

The authors would like to thank the CDISC organization, and especially the ADaM and SDS teams, for providing the standards documents we used to develop this paper and our examples. We look forward to a more detailed appendix document focused on survival and time-to-even analyses.

CONTACT INFORMATION

Your comments and questions are valued and appreciated. Authors can be reached at:

Nancy Brucken
i3 Statprobe
5430 Data Court, Suite 200
Ann Arbor, MI 48108 U.S.A.
nancy.brucken@i3statprobe.com

Sandra Minjoe
Octagon Research Solutions
585 East Swedesford Rd
Wayne, PA 19087 U.S.A.
sminjoe@octagonresearch.com

Mario Widel
Roche Molecular Systems
4300 Hacienda Dr
Pleasanton CA, 94588 U.S.A.
mario.widel@roche.com