PharmaSug2016- Paper HA03

Working with Composite Endpoints: Constructing Analysis Data

Pushpa Saranadasa, Merck & Co., Inc., Upper Gwynedd, PA

ABSTRACT

A composite endpoint in a Randomized Clinical Trial consists of multiple single endpoints that are combined in order to evaluate an investigational drug with a higher number of events than would otherwise be expected during the trial. As an example of a class of applications for composite endpoint studies are Major Adverse Cardiac event (MACE) outcome studies. As a specific example, a primary endpoint might be death due to myocardial infarction. Modern treatment and intervention in a clinical setting might make this a very rare event. Other events like fatal stroke or cardiovascular hospital admission might also be rare events of interest and of comparable interest. Combining the individual events as a composite endpoint makes important events of interest less rare in the study.

The creation and use of a composite endpoint in a clinical trial is usually justified if the individual components of the composite are clinically meaningful and of similar importance to the patient, the expected effects on each component are similar, and treatment will be beneficial.

OBJECTIVE

When analyzing composite endpoints it is the time-to-event which is modeled. Such time-to-first-event studies typically account only for the first occurrence of one of the underlying constituent variables. Nevertheless subsequent events remain of interest. Therefore, all components of a composite endpoint should be separately defined as secondary endpoints and reported with the results of the primary analysis for the purpose of relative effectiveness.

The objective of this paper is to explain how the analysis datasets can be created for the use of composite time to event analysis in such a way as to permit use of analysis ready datasets and achieve quality and efficiency while preserving traceability to main domains.

BASIC

Thorough planning, effective communication, and collaboration are essential factors for success in time-to-event studies using composite endpoints. The first thought at the time of creation of time-to-event datasets is that as programmers we know that all the documents and references related to the process are ready: Once the analysis data specifications are ready we can start creating analysis datasets. When the analysis meta data is ready we know what is to be presented. The specification of the table package for the analysis informs us as to how the study results are to be presented. All three of these document sets are essential for dataset creation.

The second thought should be to revisit the data set specification: Should the programmer go check with the statisticians and ask “Are you sure you have everything you need for near term analysis and for longer time frame analysis?” This is where it takes time and energy to keep the analysis on schedule. The programmer should brainstorm and collaborate effectively and make collective decisions with the statisticians and clinicians to gain an understanding of the sorts of questions that may arise in the future in order to create a high quality dataset with the built-in flexibility to handle requests efficiently.

This section will present the steps for creating an ADTTE dataset for a study using a hypothetical Statistical Analysis Plan.

Hypothetical Study Design

A Multicenter, Double-Blind, Randomized Study to Establish the Clinical Benefit and Safety of MK _000 vs Placebo in High-Risk Subjects Presenting With Acute Coronary Syndrome”

Primary Endpoint of Interest

Selection of these endpoints is purely for presentation purposes. There is no significant association to any clinical facts.

"Major Adverse Cardiovascular Event (MACE1) is defined as a composite of cardiovascular death, fatal myocardial infarction (MI), and fatal stroke. A myocardial infarction can be either fatal or non-fatal. Similarly a stroke can be either fatal or non-fatal." It is important to keep these facts in mind as will be seen later in the paper.

STEP 1:

Understand what need to be presented. The following figure is an example of a primary analysis table that needs to be produced for the CSR.

Figure 1: Table mock-up from the table package

Analysis of Primary Composite Endpoint:
Cardiovascular Death, Fatal Stroke, or Fatal MI
(Protocol-defined ITT Population)

	MK_000 (N=5000)			Placebo (N=6000)			Treatment Comparison	
	n (%)‡	Annual %§	KM% (95% CI)‖	n (%)‡	Annual %§	KM% (95% CI)‖	HR (95% CI)¶	p-Value¶
Primary Composite Endpoint	1200 (xx.x)	x.xx	xx.xx (xx.xx , xx.xx)	1000 (xx.xx)	x.xx	xx.xx (xx.xx , xx.xx)	0.xxx (0.xxx , 0.xxx)	0.xxx
Cardiovascular death	400 (x.xx)			200 (x.xx)				
Fatal MI	300 (x.xx)			500 (x.xx)				
Fatal stroke	500 (x.xx)			300 (x.xx)				

If a subject has multiple events only the first event contributing to the composite endpoint is counted in the component row of the table.
† xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
‡ Crude number of events (n) and percentage (%).
§ Annual% = total events / time at risk in years.
‖ Kaplan-Meier estimate and confidence interval at x years.
¶ Hazard ratio (MK_000 vs. Placebo), confidence interval, and p-Value from Cox proportional hazard model with covariates of the stratification factors xxx and treatment.
N = Number of subjects in protocol-defined ITT population.

Figure 2: Part of the Analysis metadata

Display Identifier	Result Identifier	PARAMCD	Analysis Variable	Selection Criteria/and programming statement

Table 5.1 to 5.25	Same procedure for all tables.	There are 25 parameters identified in the SAP.	AVAL,CNSR	Steps:
				1) Take CNSR=0 from ADTTE
				2) Merge with ADSL and filter for ITTFL=1
				3) Assign CNSR=1 for at risk population the working dataset is "indata"
				proc phreg data=indata;
				class xx1;
				model aval*cnsr(1)=TRTPN xx1 xx2 xx3 /risklimits alpha =0.05;
				
				proc lifetest data=indata timelist=&time outsurv=km reduceout ;
				time aval*cnsr(1);
				strata TRTPN;
				run;
				&time =estimate up to years

Looking at the documents shown in the figures and reading the SAP thoroughly we have a pretty good understanding of how to structure the ADTTE dataset for purposes of creating the analysis tables needed for the study.

The expectations in the process:

1) Traceability
2) Analysis ready
3) Easy review
4) Flexibility to add more endpoints
5) Recurrent analysis will be identified for exploratory endpoints
6) No extra effort for submission

Taking Advantage of the Intermediate Data Technique:

Intermediate datasets are created when working with complex derivations and/or smaller datasets are created using larger parent datasets. The source datasets for intermediate datasets can be a subset of SDTM or ADaM. Due to the number of endpoints and associated components and their complexity, the intermediate dataset is necessary for the process.

As mentioned above when analyzing composite endpoints it is necessary to analyze each component associated with those composites. Composite endpoints can be described as a function of more than one component. To reiterate, each component is a function of other information collected in the CRF. For example, a primary composite endpoint can be a combination of stroke, myocardial infarction and cardiovascular death . Additionally, in the CRF stroke can be collected as either fatal or non-fatal. Obviously then, when doing analysis the definition for each component is very important.

STEP 2: Preprocessed data

In this example the required information has been collected in CF Domain, an example of which is in Table 1.

Table 1: Part of original CF Domain

CFSPID	CFTESTCD	CFTEST	CFORRES	SUBJID	cfseq
19050	DEATH	Death	No	51015	268695486691332
19050	INVREP	Reported Event	Chest tightness	51015	268695486691334
19050	ONSETDTC	Event Onset	2009-05-01	51015	268695486691335
19047	DEATH	Death	No	51078	268733183262332
19047	INVREP	Reported Event	worsening angina pectoris	51078	268733183262334
19047	ONSETDTC	Event Onset	2009-04-17	51078	268733183262335
19028	CARDEVNT	Cardiac Event	CORONARY	51093	270641683796379
19028	DEATH	Death	No	51093	270641683796373
19028	INVREP	Reported Event	Worsening Ischaemic heart disease	51093	270641683796375
19028	ONSETDTC	Event Onset	2009-01-28	51093	270641683796376
19078	DEATH	Death	No	51114	270493958947336
19078	INVREP	Reported Event	Worsening coronary artery disease	51114	270493958947338
19078	ONSETDTC	Event Onset	2009-04-30	51114	270493958947339
19002	DEATH	Death	No	51117	266212612948391
19002	INVREP	Reported Event	Cerebrovascular accident	51117	266212612948393
19002	ONSETDTC	Event Onset	2008-08-25	51117	266212612948394

When you look at the CFTEST and CFORRES variables you are able to see that all the information needed for the components are collected in the CFORRES variable. This makes data manipulation tedious. When the data was processed in the data repository a special id named CFSPID was added to identify each unique collection. In this kind of situation that variable serves a useful benefit to the programmer.

Looking at the data it is obvious that missing date imputation should be performed at an intermediate level since it makes the ADTTE creation process easier. Another reason is that there may be many instances in which we may have to merge with another domain in order to get accurate information. For example, one important component of MACE studies is "Unstable Angina." In order to get the correct event date we may need to merge with hospital data and compare the reported occurrence date to be sure that it is consistent with the hospital admission date. In such instances we need to merge the CF data and hospital data. The efficient way to handle all of these cases is to have a clean dataset that can be easily manipulated. The best approach is to have a horizontal intermediate dataset. The horizontal structure of the intermediate dataset makes manipulation easier.

In our example all of the patients who have experienced at least one event are included in the intermediate dataset. The method for creating the horizontal data set is illustrated with the following simple code and the result is shown in Table 2.

```
proc transpose data=CF out=TCF
var cforres;
id cftestcd;
idlabel cftestcd;
by usubjid cfspid
run;
```

Table 2: Transposed Data set

SUBJID	CFSPID	DEATH	INVREP	ONSETDTC	CARDEVNT	ACIDTL
51015	19050	No	Chest tightness	5/1/2009		
51078	19047	No	worsening angina pectoris	4/17/2009		
51093	19028	No	Worsening Ischaemic heart disease	1/28/2009	CORONARY	
51114	19078	No	Worsening coronary artery disease	4/30/2009		
51117	19002	No	Cerebrovascular accident	8/25/2008		
51147	19298	No	Chronic Subdural Hematoma	9/5/2012		
51153	19025	No	Small myocardial infarction	1/30/2009	Acute Myocardial Infarction	Fatal MI
51189	19256	No	Acute Cardiac Failure	2/6/2012	CONGESTIVE HEART FAILURE	
51189	19265	No	Acute Cardiac Failure	4/7/2012	CONGESTIVE HEART FAILURE	
51234	19251	Yes	Sudden Death	12/25/2011	DEATH	
51261	16160	No	Ascending aortic dissection	5/9/2010		
51276	19026	No	Vasovagal reaction	1/6/2009		
51276	19029	No	Confusion	2/11/2009		
51360	19031	No	Stroke	1/22/2009		
51366	19217	No	unstable angina pectoris	2/6/2011	UNSTABLE ANGINA PECTORIS	
51381	19064	No	Pulmonary embolism	10/4/2008		
51423	19260	No	Myocardial Infarction	2/27/2012	Acute Myocardial Infarction	Non Fatl MI

STEP 3: Creating composite endpoints for ADTTE dataset

Using the horizontal data we now create new variables to replace the missing dates. Other related issues should be taken care of as well since the "AVAL" is created using the date. Additionally, two important derivations should be taken care of at this point.

a) If two events have occurred on the same day then there should be a hierarchy for picking one of them as the primary event. For example, if death and myocardial infarction have occurred on the same day for a given patient then which one will be considered the first event when one of them must receive the designation as 'first?' Usually this hierarchical order is decided by clinical team members and we should have a variable for this in the data for the purpose of sorting by order.

b) The dataset should support analysis beyond the requisite inferences. Often we need to find the number count for each composite endpoint and counts for events within the components. For example how many patients are included in the primary endpoints? Out of those, what are the counts for each category?
Creating composite endpoints:

```
data mace1; PARAMN=1;
 set adevent;
     if CARDEVNT="DEATH' and Death ='Yes' then do;
       adt=onsetdti;
      COMPO=CARDEVNT                          ←-First component of Primary endpoint ;
         output;
        end;
      if CARDEVNT ="Acute Myocardiac Infacrction' and ACLIDTL="Fatal MI" then do;
            adt=onsettdt;
          COMPO=ACLIDTL;                      ←-Second component of Primary endpoint
          output;
         end;

   (add component defined in the protocol )   ←-Third component of Primary endpoint

    format adt date9.;
    keep subjid adt paramcd compo ;
    run;
```

COMPO variable is important. The individual counts should be displayed in the analysis table. We can get total counts for composite endpoint and each component by treatment group using one freq statement.

Censor date: Algorithms for censored dates, survival date, or last contact date for MACE outcome studies are complex and need careful data manipulation, and the criteria are explained in the SAP. Creating censored dates is outside of the scope of this presentation. The censored date would be included in the ADSL dataset.

STEP 4: Creating the final ADTTE dataset involves the following sub-steps:

a) Merge and get required variables from ADSL,
b) Create AVAL
c) Identify events which fall between start date and survival date
d) Sort by usubjid paramcd adt and assign CNSR for observations identify in c)

Table 3: Final ADTTE Dataset

SUBJID	COMPO	EVNTDESC	PARAMCD	SCRDOM	CNSR	STARTDT	ADT	AVAL	INCLUDE
51015	Fatal MI	CV Death/Fatal MI/Fatal Stroke	MACE1	ADEVENT	0	1-Mar-09	9-Jun-10	466	Y
51015	Fatal Stroke	Fatal MI/Fatal Stroke	MACE2	ADEVENT	0	1-Mar-09	9-Jun-10	466	Y
51279	CV Death	CV Death/Fatal MI/Fatal Stroke	MACE1	ADEVENT	0	14-May-09	12-Apr-10	334	Y
51279	Fatal MI	Fatal MI/Fatal Stroke	MACE2	ADEVENT	0	14-May-09	12-Apr-10	334	Y
51300	Fatal Stroke	CV Death/Fatal MI/Fatal Stroke	MACE1	ADEVENT	0	14-Jun-09	21-Jul-09	38	Y
51300	Fatal MI	CV Death/Fatal MI/Fatal Stroke	MACE1	ADEVENT		14-Jun-09	8-Mar-10	268	Y
51300	Fatal MI	Fatal MI/Fatal Stroke	MACE2	ADEVENT	0	14-Jun-09	21-Jul-09	38	Y
51300	Non-fatal MI	Fatal MI/Fatal Stroke	MACE2	ADEVENT		14-Jun-09	8-Mar-10	268	Y
51556	CV Death	CV Death/Fatal MI/Fatal Stroke	MACE1	ADEVENT	0	8-Jul-09	2-Mar-11	603	Y
51556	Fatal MI	CV Death/Fatal MI/Fatal Stroke	MACE1	ADEVENT		8-Jul-09	8-Feb-13	1312	
51556	Fatal MI	Fatal MI/Fatal Stroke	MACE2	ADEVENT	0	8-Jul-09	10-Mar-11	611	Y
51556	Fatal- MI	Fatal MI/Fatal Stroke	MACE2	ADEVENT		8-Jul-09	16-Feb-13	1320	

In these sub-steps and in the final data shown above AVAL is the elapsed time to the event of interest or censoring from the origin. PARAM describes the contents of AVAL. STARTDT is the date of origin from which AVAL is calculated. STARTDT can be the equivalent of the randomization date, or treatment start date. ADT represents the analysis date for the event or censoring associated with AVAL.CNSBR=0 and indicates an event in baseline follow up is indicated in CNSR. The reasons for censoring are contained in EVNTDESC and indicates the event of interest or the event that warrants censoring

Assessing the ADTTE creation for composite endpoints:

Traceability: Traceability is one of the most important features of the ADTTE dataset. Traceability is important when evaluating from where the results came. In this case, we need to identify the source of the analysis date that is our intermediate dataset ADEVENT. When an intermediate dataset is created with a horizontal structure the CFSEQ information related to onset date, which eventually will become ADT, was lost. But the CFSEQ was merged back to intermediate ADEVENT dataset using the same code but with the different value for VAR option. With the use of the idlabel option it is easier to track which variable should be used to trace the sequence for onset date.

Similarly, data point traceability is implemented in a stepwise fashion. SRCSEQ in ADTTE points the analyst to the particular record in ADEVENT from which ADTTE.ADT was selected. SRCDOM, SRCSEQ, and SRCVAR in ADEVENT will in turn point to the precise source from which ADTTE.ADT was ultimately obtained. By using data point traceability, the location and existence of all dates can be systematically verified.

The metadata/specifications become dramatically simpler, as the derivations can be separated into individual components. Variables in ADaM datasets are often dependent on the associated parameters. For example, ADT may have one derivation for a particular value of PARAM, but ADT may have a different derivation for another value of PARAM.

Analysis Ready:

ADTTE should be “analysis-ready,” meaning it should contain all of the variables needed for the specific analysis, so that the analysis can be replicated by performing the actual statistical test without first having to manipulate the data. Looking at the final dataset you can see all the information we need for the analysis is included in the ADTTE dataset. One exception is information about at risk patients.

After the risk benefit analysis we need to make a decision about whether it is efficient to add “at risk” patients to the final ADTTE dataset. Should the censored patients who never had any events be included? In the example being used here two risks were identified. To include those cases we need to subset records using CNSR=0 and get a smaller dataset which has unique patients and unique parameters. In the next step we then need to create a complete dataset by merging with the population of interest. Having a complete dataset is good, but a potential risk arises if there are large numbers of end points and a large number of patients. In that case the analysis dataset can exceed the 1 GB that is the allowed size for a submission. The other risk is that if we remove all other observations except those for which CNSR=0 then it can lead to another issue. That risk is that it is possible that

recurrent analyses normally performed for exploratory purposes for MACE studies will not be able to use the same data set.

Currently the only manipulation needed to be done for the analysis is to merge the subset of the data using the criteria in the "analysis meta data" with the desired population in ADSL and assign CNSR=1. We are taking the survival date /last contact from the ADSL dataset for patients who never had any events. This is certainly more efficient as survival date/last contact date in ADSL's metadata is already traced to the source data.

Exploratory Analysis: Any exploratory analysis can be performed using the final ADTTE dataset since it preserves all events that occurred during the duration of the study.

Submission ready: This data is submission ready. All requirements that are needed for submission data are contained in the final data. Once the decision is made as to which endpoints will be included in the Analysis Data Review aid package the only data manipulation that needs to be done is to subset the data for the desired parameters.

Data Review/Validation: There are significant advantages to having a predecessor dataset, ADEVENT, with a horizontal structure since it gives significant advantages to the statistician and/or reviewer. The transposed dataset can be easily examined to determine which date is selecting for ADTTE data and why it has been selected .

Conclusion: The example presented in this paper illustrates the method for creating a time-to-event dataset for composite endpoints in cardiovascular studies. MACE studies involve substantial statistical analyses using information collected for years for a large number of patients. In such large studies the methods illustrated here highlight the use of intermediate datasets for traceability. This data creation process significantly aids any sensitivity analysis that will be performed at any time. One of the most important elements of the process is creating intermediate datasets and taking care of missing or questionable dates at an early stage. This makes subsequent manipulations necessary for constructing the ADTTE far easier with the added advantage of more straightforward traceability. Horizontal structure of the intermediate dataset allow us to add endpoints to ADTTE data anytime without having touched the dataset. It serves as a valuable data repository .
Finally data structure of ADTTE can be easily used for any study which has more than one phase , The only difference is adding more indicator variables to identify the phases and creating censor variables for each phase.

ACKNOWLEDGMENTS

Thanks to Changhong Shi for continuous support and encouragement .

CONTACT INFORMATION

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

Pushpa Saranadasa
351 North Sumneytown Pike
North Wales
PA 19454
Email :Pushpa_saranadasa@merck.com

TRADEMARK