Creating Macros for Survival Data in Oncology Study

Jagannath Ghosh, MedFocus LLC, Chicago, Illinois

Abstract: In this paper, we introduce some system functions and macros to create study specific survival variables for oncology clinical trials. We present five variables which show overall survival time, time to disease progression, duration of response, progression free survival and time to treatment failure for one particular study and then we will be using survival analysis based on overall survival and censoring information. Note that the study data will be fake data, however, it is still valuable for a real life data generated from a clinical trial. The purpose of his paper is to show the power and usefulness of SAS in clinical research (basically studies which require death and survival information, such as cancer and HIV).

Introduction: In oncology study, we most often perform survival analysis for the efficacy measurement. Most frequently used efficacy measurements are overall survival (OS), and progression free survival (PFS). Duration of response (RPDUR), time to treatment failure (TTTF) and time to disease progression (TTPD) are also widely used efficacy measurements. Using our macros, we will be creating those efficacy variables and then will limit our discussion to overall survival.

Variable Definition: Clinical trial starts with the protocol development. Every study is unique based on hypothesis we need to address, however, data structure might have some similarity. Overall survival, time to disease progression, duration of response, progression free survival and time to treatment failure all are based on duration. Now the concern is how to calculate those durations. Usually when patients are enrolled in a study, we do have either randomization record if the study has more than one arm or first dose administered date or study start date. We pick one of those dates according to protocol to calculate the efficacy variables described earlier. The overall survival time is defined as

OS= (last contact date-randomization date+1)/(365.25/12) in months.

Last contact date will be found either in summary data set or death data set or follow up data set or in all of them. For simplicity, those are the sources from where the dates will be collected. Based on those dates, the last contact date for each patient is determined [**Table 1**] and the censor information is created if the patient is still alive or got lost.

Time to disease progression is calculated based on the criteria of disease progression. Usually, the RECIST criteria are used to determine the disease progression dates. The disease progression time [**Table 1**] will be calculated based on the formula set below:

Progression time = (First progression date-randomization date+1)/(365.25/12) in months

In oncology, response is defined as lesion's disappearance completely or partially. Duration of response (RPDUR) [**Table 1**] is measured between the best overall response

and the earliest of progressive disease date or death date. Best overall response is the best response recorded from the start of treatment until disease progression or recurrence. In general patient's best response will depend on the achievement of measurement criteria (RESIST) [3].

Duration of Response= (earliest of (PDDT, DTHDT) – best overall response+ 1) / 30.4375 in months

Other two important efficacy measurements are progression free survival and time to treatment failure. Progression free survival can be calculated from the earliest date of disease progression or death date to randomization date while time to treatment failure is measured from the earliest of study termination date, progression disease date and death date to randomization date. The end date is usually study specific and is defined by the protocol.

Progression free survival= ((earliest of (PDDT, DTHDT) - randomization date + 1) / 30.4375 in months

Time to treatment failure = (min ((ENDDT, PDDT, DTHDT) – randomization date + 1) / 30.4375 in months

The above formulas indicate that death date, disease progression date, last contact date, study end date and best overall response are the key for oncology efficacy study. To determine all those dates, data structure of the study needs to be discussed. In real data, progression date can be recorded [**flow chart 1**] in follow up dataset, summary dataset or response dataset. The conservative way to determine first progression date is to stack all the progression disease dates from various data sources and then select the earliest date as the disease progression date. Similar technique can be used to determine the death date and the best overall response date. Study ends date can be found by the last date of the study participation. Note that censoring information is not included in this paper; however, in coding the FDA guideline is followed.

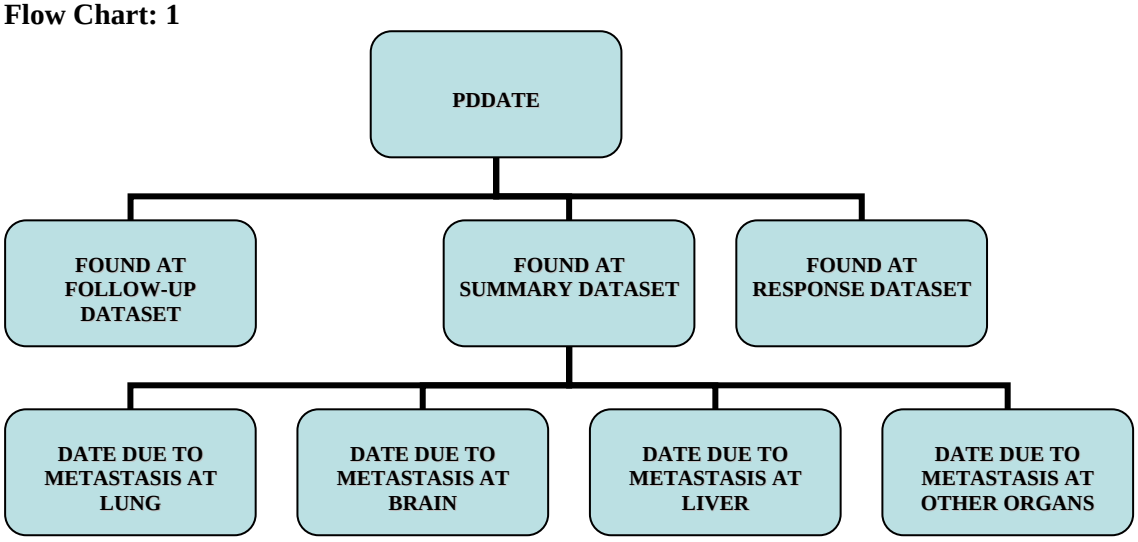
SAS Macros: In **flowchart 1**, it is noted that disease progression can be metastasis. To distinguish those metastasis differences, different variables might be created. For example if disease progress is found in brain, the variable can be created as 'braindta' or if it is found in lung, it can be named as 'lungdta' but all represent the disease progression date. In our first macro '%GETDATE' stacks those variables according to condition(s) described in study protocol and finally it assigns new data set's name for future use.

```
%GETDATE(libref=fake,summary=summary,condiVar=DISEASE,Condi=1,sortlist=PATIENT,
varlist='LUNGDTA LIVERDTA BRAINDTA OTHERDTA',assignName=summary211);
```

In %GETDATE macro, variable(s) list can vary depending on the data structure. To accommodate varying variables, the following piece of code is used to create macro variables so that stacking can be easier.

```
data _null_;
whereisdate=0;
i=0;
do until(whereisdate=0);
whereisdate=findc(&varlist,' ',whereisdate+1);
if whereisdate=0 then put "The End";
else do;
i=i+1;
p=length(scan(&varlist,i+1));
whatfound=substr(&varlist,whereisdate+1, p);
%let found=whatfound;
call symput('newvar'||left(put(i,2.)),&found);
call symput('nvar',i);
end;
end;
run;
```

Flow Chart: 1



The above piece of code is the main coding to derive % GETDATE macro. This macro will produce four data sets internally since varlist (above) has four variables and finally it will stack them and will assign a name as summary_r.

```
data &assignName;
set &assignName  mm&j(rename=(&&newvar&j=&fstVar));
if &fstVar ne .;
run;
```

Note that *findc* function is not a part of SAS 8.2 and thus this macro will not work if we do not use SAS 9.0 or later.

Progression disease date is also recorded in follow-up and response data set. %GETDATE macro captures progressive disease date from different data sets and stores them in different assigned data sets. Later another macro named %INDICTDATE uses those data sets.

```
%INDICTDATE (datasetlist='Response summary_r summary_r_f summary_r_s',
sortlist=Patient, pickObs=Patient,wantName=PDDATE, creatVar=PDDT, whichObs=first);
```

In %INDICTDATE macro dataset list indicates where the disease progression dates are found. After stacking them, the first observation is desired to get the first disease progression date. The underline code is used to pick the first disease progression date:

```
Data &wantName;
 set &wantName;
 by &sortlist;
 if upcase(&whichobs)='FIRST' then do;
 if first.&sortlist;
 end;
 else do;
 if last.&sortlist;
 end;
 drop &whichObs;
 label &creatVar="&wantName";
 run;
```

In %INDICTDATE macro the parameter 'datasetlist' requires all those data sets Created by %GETDATE macro from which first progressive disease (PD) date is picked. Similar techniques are applied to pick the death date and best overall response date, however, last study date is chosen by the last known date of study participation.

Finally, the macro %SURVDATA creates final survival data set according to protocol and the formulas mentioned above.

```
%SURVDATA(libref=fake, sortlist=patient, randdata=txinduct, randdate=txdta,
                dthdata=deathdate,lcontactdata=lastcontact,pddata=pddate);
```

%SURVDATA macro uses data sets that are created earlier (using %INDICTDATE).Inside this macro, the following code is written to find the time to disease progression and its censoring. Similar technique is used to find the OS, PFS, TTTF and DOR. Finally the name of the survival data set created by this macro is **SurvData [Table 1].**

```
min_lc_death=min(lstdt,dthdt);
format min_lc_death date9.;
if pddt ne . then do;
ST=(pddt -txdta +1)/365.25*12;
censor=0;
STCD='TTPD';
end;
else do;
ST=(min_lc_death-txdta+1)/365.25*12;
censor=1;
STCD='TTPD';
end;
```

Table 1. Derived Survival data

PATIENT	TXDTA	DEATHDATE	LASTCONTACT	PDDATE	ENDDATE	BOR	STCD	censor	ST
03-3002	09OCT2002	.	21APR2003	.	.	15NOV2002	OST	1	6.41
03-3002	09OCT2002	.	21APR2003	.	.	15NOV2002	PFS	1	6.41
03-3002	09OCT2002	.	21APR2003	.	.	15NOV2002	RPDUR	1	5.19
03-3002	09OCT2002	.	21APR2003	.	.	15NOV2002	TTPD	1	6.41
03-3002	09OCT2002	.	21APR2003	.	.	15NOV2002	TTTF	1	6.41
04-4015	03NOV2004	.	05DEC2005	14SEP2005	.	28MAR2005	OST	1	13.08
04-4015	03NOV2004	.	05DEC2005	14SEP2005	.	28MAR2005	PFS	0	10.38
04-4015	03NOV2004	.	05DEC2005	14SEP2005	.	28MAR2005	RPDUR	0	5.62
04-4015	03NOV2004	.	05DEC2005	14SEP2005	.	28MAR2005	TTPD	0	10.38
04-4015	03NOV2004	.	05DEC2005	14SEP2005	.	28MAR2005	TTTF	0	10.38
04-4016	16DEC2004	.	10NOV2005	14JAN2005	30DEC2004	.	OST	1	10.84
04-4016	16DEC2004	.	10NOV2005	14JAN2005	30DEC2004	.	PFS	0	0.99
04-4016	16DEC2004	.	10NOV2005	14JAN2005	30DEC2004	.	TTPD	0	0.99
04-4016	16DEC2004	.	10NOV2005	14JAN2005	30DEC2004	.	TTTF	0	0.49
04-4017	22DEC2004	11FEB2006	11FEB2006	08APR2005	18APR2005	.	OST	0	13.70
04-4017	22DEC2004	11FEB2006	11FEB2006	08APR2005	18APR2005	.	PFS	0	3.55
04-4017	22DEC2004	11FEB2006	11FEB2006	08APR2005	18APR2005	.	TTPD	0	3.55
04-4017	22DEC2004	11FEB2006	11FEB2006	08APR2005	18APR2005	.	TTTF	0	3.55
04-4018	03JAN2005	11FEB2006	20DEC2005	04NOV2005	08NOV2005	.	OST	0	13.31
04-4018	03JAN2005	11FEB2006	20DEC2005	04NOV2005	08NOV2005	.	PFS	0	10.05
04-4018	03JAN2005	11FEB2006	20DEC2005	04NOV2005	08NOV2005	.	TTPD	0	10.05
04-4018	03JAN2005	11FEB2006	20DEC2005	04NOV2005	08NOV2005	.	TTTF	0	10.05
04-4019	15JUN2005	.	25JUL2006	27MAR2006	.	18OCT2005	OST	1	13.34
04-4019	15JUN2005	.	25JUL2006	27MAR2006	.	18OCT2005	PFS	0	9.40
04-4019	15JUN2005	.	25JUL2006	27MAR2006	.	18OCT2005	RPDUR	0	5.29
04-4019	15JUN2005	.	25JUL2006	27MAR2006	.	18OCT2005	TTPD	0	9.40
04-4019	15JUN2005	.	25JUL2006	27MAR2006	.	18OCT2005	TTTF	0	9.40

ST= Survival Time, BOR=Best Overall Response, STCD= Survival Time Coding

Survival Analysis: Once the survival data set is derived, what remains is how to analyze this data. For survival data, we use PROC LIFETEST which needs duration of time variables and censoring information. The following code can be used to analyze OS.

```
data os;
set paper;
if stcd='OST';
run;
ods output productlimitestimates=surv
                    quartiles=surv_quartiles;
proc lifetest data=os;
time ST*censor(1);
run;
```

The output of survival probability and survival quartiles are described in **Table 2** and **Table 3** respectively.

Table 2: Survival Probability

Survival Time	Survival Probability	Failure	Standard Error	Number Failed	Number Left
0.00	1.00	0	0	0	6
6.41	.	.	.	0	5
10.84	.	.	.	0	4
13.08	.	.	.	0	3
13.31	0.67	0.3333	0.2722	1	2
13.34	.	.	.	1	1
13.70	0.00	1.0000	0	2	0

Table 3 is more frequently used than Table 2 because median survival time is one of the pivotal parts in report writing for oncology research.

Table 3: Quartile Estimates of Survival Time

Percent	Point Estimate	Lower 95% Confidence Limit	Upper 95% Confidence Limit
75.00	13.70	13.31	13.70
50.00	13.70	13.31	13.70
25.00	13.31	13.31	13.70

Discussion: In this development process, we ignored the fact that dates variables can be stored in character form as well as numeric. It is likely to have some missing information in either of these two forms. Missing values can be handled using another macro named '%MISSINGDATE' [1]. Using this macro numerical missing dates are replaced by the character dates in case character dates exist. However, no numeric dates will be replaced for character dates because calculation for survival times is based on numeric data.

The ordering of the variables in every data set is very important since the macros stack the date variable after the sort variables. Without proper order the macros will produce the anomalous results.

Those macros will fail to function in case several conditional variables are used. If there exit several conditional variables, %GETDATE macro can be called for each variable.

The system functions are used to convey the message to users if the dataset and library do not exit. Those messages will be written in the log function and output will not be produced. Using system functions, those macros can be made for automated results. In this situation, data structure needs to be well defined.

Conclusion: Three macros have been created because of the variation of the data from study after study. In most oncology study variables and style varies for each study. To accommodate almost every study, three instead of one macro have been implemented, however, if data management does follow the common structure across the studies, three macros can easily be combined.

References

1. Ghosh, J. "Macro for Managing Date Variable(s) in Oncology Research", *Western User of SAS software (WUSS)* 2008 proceedings

2. SAS Institute Inc. (2004), SAS Online Doc, Version 9.0, Cary, NC

3. FDA Guidance for Industry. "Clinical Trail Endpoints for the Approval of Cancer Drug and Biologics", http://www.fda.gov/cber/guidelines.htm : April 25,2005

Contact Information:

Please send your comments and questions to:
Jagannath Ghosh
MedFocus LLC
Chicago, Illinois
Contact phone: 530-220-9066
Email: ja_ghosh@yahoo.com