Surviving Left Truncation Using PROC PHREG

Aimee J. Foreman, Ginny P. Lai, Dave P. Miller,
ICON Clinical Research, San Francisco, CA

ABSTRACT

Many of us are familiar with PROC LIFETEST as a tool for producing survival function estimates, but it is not widely known that PROC PHREG can also be used for this purpose. This is especially convenient in the case of left truncated data. Left truncation is present, for example, in studies of disease mortality where survival from the time of diagnosis is the outcome of interest even though patients may have been diagnosed many months or years prior to enrollment in the study. While PROC LIFETEST is not set up to handle this situation, PROC PHREG is, using the ENTRY= option to specify the left truncation time. Group differences in survival can be estimated with STRATA processing with a null model, which avoids the assumption of proportional hazards. We will provide examples from a disease registry that show why the issue of left truncation is important when analyzing survival data, and how ignoring it can get you into trouble.

INTRODUCTION

Survival curves are commonly created to examine the time until a group of subjects experiences an event such as death. Often used in the preliminary stages of survival analyses, they provide more information about overall risk than individual statistics such as the median or overall percent survival. At each time point (x-axis value) on a survival curve, the survival function estimate (y-axis value) represents the estimated probability of surviving beyond that time. For typical time-to-event data, where all subjects are followed from a natural starting point until they experience an event or are right-censored, survival estimates are easily obtained from PROC LIFETEST. For left truncated data however, where some subjects do not enter the risk set until a known period after the time origin, special methods must be used that are not available in PROC LIFETEST.

Left truncation often arises when patient information, such as time of diagnosis, is gathered retrospectively. For example, in a study of disease mortality where the outcome of interest is survival from the time of diagnosis, many patients may not have been enrolled in the study until several months or years after their diagnosis. Those patients, by virtue of having survived to the time of enrollment, could not have had an event between diagnosis and study enrollment, and therefore they should be removed from the risk set between those two time points. To leave them in the risk set would bias the survival estimates.

Though PROC PHREG offers two methods for accommodating left truncation, it is typically used to perform Cox proportional hazards regression analysis and not for plotting survival curves. However, using code that is straightforward and simple PROC PHREG can output Kaplan Meier estimates of survival that are equivalent to those estimated by LIFETEST when the data are not left truncated. With the addition of the ENTRY= option in the MODEL statement, we will provide examples of how to similarly output survival function estimates that accommodate left truncated data. Since survival curves are often used to compare mortality risk across different groups of patients, we will also show how to adapt this code for stratified survival estimates.

A basic understanding of survival analysis is assumed in this paper, as is familiarity with SAS/STAT® software. Several alternative coding methods will be suggested in the final section.

CREATING SIMPLE SURVIVAL CURVES

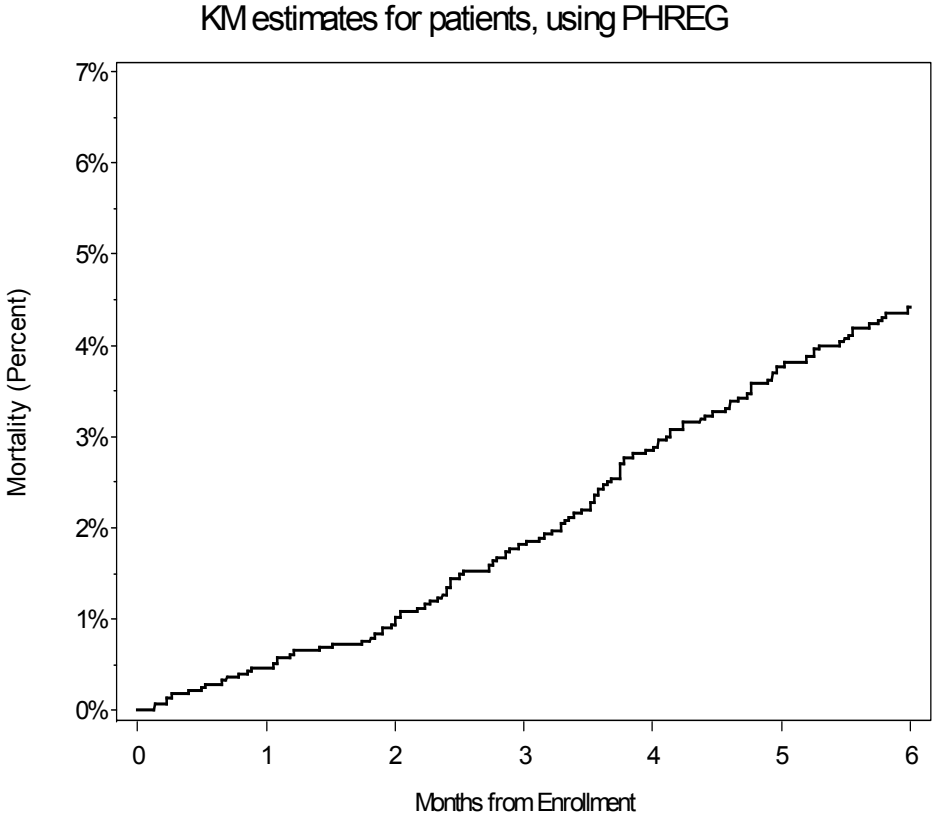
PROC LIFETEST

Most SAS users are familiar with obtaining Kaplan Meier (KM) survival estimates and plotting survival curves using PROC LIFETEST. The following code shows the example of generating KM estimates from simple time-to-event data. Here, **tt_bl_dth** represents time from baseline (enrollment) to the event or right-censoring. The variable **death** is a censoring variable that equals 1 when a patient has died, and values of 0 are censored.

```
ods listing close;
ods output productlimitestimates=temp02 (keep= tt_bl_dth failure survival censor
                                          where=(failure ne . or censor eq .)) ;
proc lifetest data=temp01;
  time tt_bl_dth*death(0);
run;
ods listing;
```

The output data set temp02 is used to plot the curves. For all examples in this paper, we will plot mortality (1 – survival) probability.

Figure 1.



PROC PHREG

Although LIFETEST is more typically used to get KM estimates of survival, we can also achieve the same goal using PROC PHREG. This became possible after release 8.1 which allowed for a null model to be specified in the MODEL statement. The survival estimates (S) from the output data set temp03 are nearly identical to the estimates we obtained from PROC LIFETEST.

```
proc phreg data=temp01;
  model tt_bl_dth*death(0)=;
  output out=temp03 survival=S;
run;
```

Two additional steps need to be taken in order to create the mortality curve. First, as PROC PHREG does not generate failure time (1-S) automatically from the OUTPUT statement, this must be calculated in the DATA step. And second, the output data set temp03 does not include an estimate at time=0, so we must write out an observation in the DATA step where failure=0 and time t=0. The following code demonstrates how this is done:

```
proc sort data=temp03 out=temp03_s;
  by tt_bl_dth;
run;

data temp03_s;
  if _n_ eq 1 then do;
    t = 0;
    failure=0;
    output;
  end;
  set temp03_s;
  t=tt_bl_dth;
  failure=1-S;
  output;
run;
```

LEFT TRUNCATED DATA

To illustrate the issue of left truncated data in survival analysis we will use a registry of approximately 3000 patients who were enrolled in the study regardless of whether they were newly diagnosed with a particular disease, or were diagnosed months or years previously. About two years of follow-up is available from the time of enrollment, while there are patients with up to 40 years of "follow-up" from the time of diagnosis. Of course, patients with such extensive survival data will not enter the risk set until they enrolled in the study, so their "follow-up" is limited to the time of diagnosis, when certain tests were administered and patient information recorded, and the time from enrollment until death or discontinuation from the study.

However, this highlights an important patient stratification that is at the heart of the reason that left truncation cannot be ignored. Patients who were previously diagnosed and who died before the study began are not represented in this data set; rather, the previously diagnosed patients for whom we have data are those who survived and represent a healthier sample because they have survived. Patients who were diagnosed at the time of being enrolled, on the other hand, may include some people whose disease is advanced and will die soon after entering the study. Therefore if we were to examine survival from the time of enrollment, we would observe a survival bias.

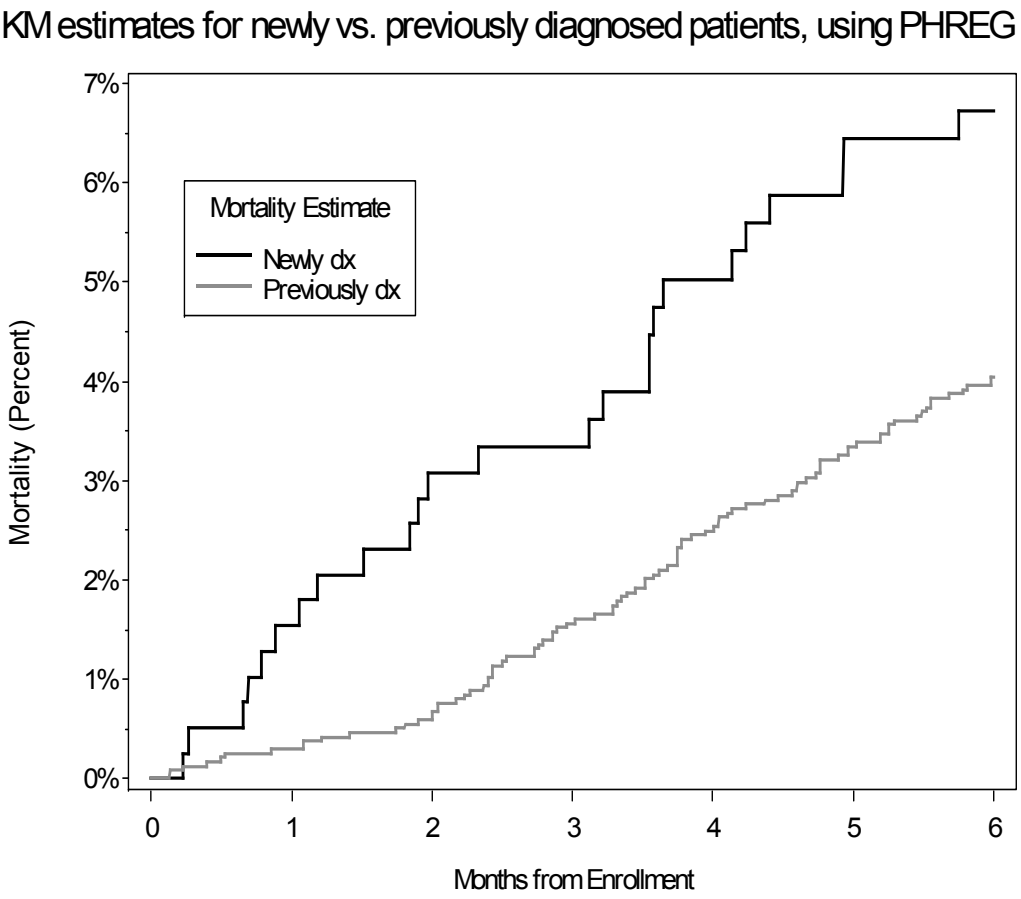
The code and figure below help visualize this phenomenon. Like the previous example, the null model is specified in PROC PHREG but now we will stratify by the variable **newdiag**. This variable has two possible values for newly or previously diagnosed patients, and the STRATA statement estimates the survival function separately for each group. Again, the survival estimates that are output do not include estimates at time 0, so to improve the look of the figure we output records at time t=0 for both strata, setting failure=0.

```
proc phreg data=temp01;
  strata newdiag;
  model tt_bl_dth*death(0)= ;
  output out=temp04 survival=S;
run;
proc sort data=temp04 out=temp04_s;
  by newdiag tt_bl_dth ;
run;

data temp04_s;
  set temp04_s;
  by newdiag;
  if first.newdiag then do;
    t = 0;
    failure=0;
    output;
  end;
  t=tt_bl_dth;
  failure=1-S;
  output;
run;
```

The resulting mortality curve, seen below in figure 2, shows an apparent difference in mortality for patients who are newly versus previously diagnosed. The survival curve for the previously diagnosed patients is smoother with fewer steps than the newly diagnosed curve because there are more patients (about 80% of the cohort) in the previously diagnosed group.

Figure 2



This difference is confirmed by the log-rank tests (below), obtained from PROC LIFETEST.

Log-rank test for Newly vs. Previous diagnosis

Obs	Test	ChiSq	DF	Prob ChiSq
1	Log-Rank	5.6300	1	0.0177
2	Wilcoxon	5.9477	1	0.0147
3	-2Log(LR)	6.6497	1	0.0099

ESTIMATING SURVIVAL FOR LEFT TRUNCATED DATA

There are undoubtedly several ways to obtain survival estimates that appropriately accommodate left truncated data. Two relatively simple options are detailed below.

THE MAYO CLINIC RESEARCH MACRO

The Division of Biostatistics at the Mayo Clinic maintains a very useful web site with locally written SAS macros (see http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm). The software provided on this site is free and downloadable. One of the macros developed by the Mayo Clinic biostatisticians, called SURVTD, produces KM survival estimates for left truncated data. It can output overall survival estimates as well as estimates stratified by risk groups.

PROC PHREG: THE ENTRY= OPTION

A more familiar alternative is PROC PHREG. As illustrated above, PROC PHREG is a suitable substitute for PROC LIFETEST for obtaining KM survival function estimates (though it lacks some of the handy features of LIFETEST, including log-rank tests). Unlike LIFETEST, it includes options that accommodate left truncation. According to the SAS documentation, the ENTRY= option to the MODEL statement will output survival function estimates for left truncated data, which we can then use to plot survival curves (an alternative method will be discussed later). This

option takes as its argument a variable with the left truncation time. In the example of the disease registry, the left truncation time is the time from diagnosis to enrollment for previously diagnosed patients.

The PROC PHREG/ENTRY= option has the advantage of being relatively transparent, whereas the Mayo Clinic macro was written to accomplish much more than outputting survival estimates and is therefore more complicated. A brief inquiry sent to Dr. Terry M. Therneau, one of the authors of the macro, confirmed that the estimates obtained from these two methods are equivalent.

OVERALL SURVIVAL

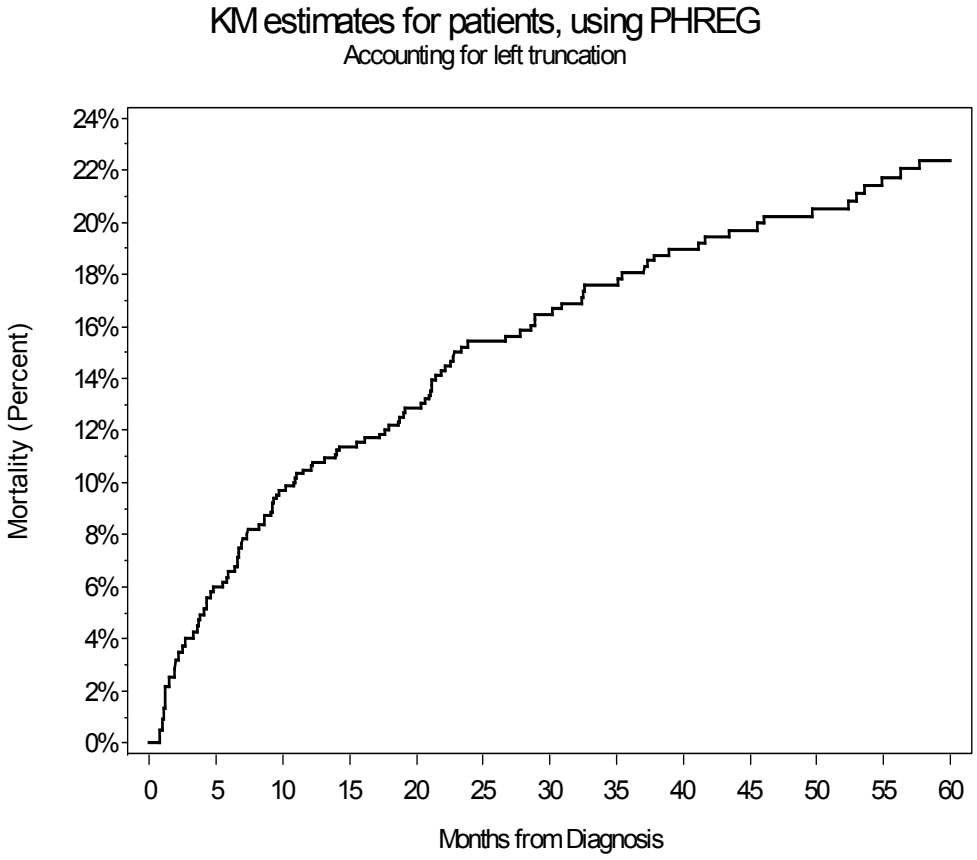
Adopting the PHREG method of accounting for left truncation, the next step is to examine overall survival from the time of diagnosis. The following code will output survival function estimates using the time from diagnosis until death or censoring (**tt_dx_dth**), where the variable **tt_dx_bl** contains the left truncation time from diagnosis to enrollment.

```
***data step code to create new censoring time from diagnosis;
tt_dx_dth = tt_dth_bl + tt_dx_bl;

proc phreg data=temp01;
  model tt_dx_dth*death(0)= /entry=tt_dx_bl;
  output out=temp06 survival=S;
run;
```

Transforming the survival estimates into mortality estimates, we plot overall mortality from the time of diagnosis in figure 3. Substantially more follow-up time is available because we are examining the time from diagnosis. This mortality curve is somewhat steeper at the start than the plot of survival from enrollment in figure 1, and this is likely due to the initial contribution of newly diagnosed patients with higher risk than previously diagnosed patients.

Figure 3



THE PROPORTIONAL HAZARDS ASSUMPTION

As we saw above with the comparison of mortality for newly vs. previously diagnosed patients, it is often desirable to plot survival curves for two or more risk groups. For patients in our study, higher values of one particular test measured at the time of diagnosis are believed to be predictive of mortality, so we would like to check this by plotting

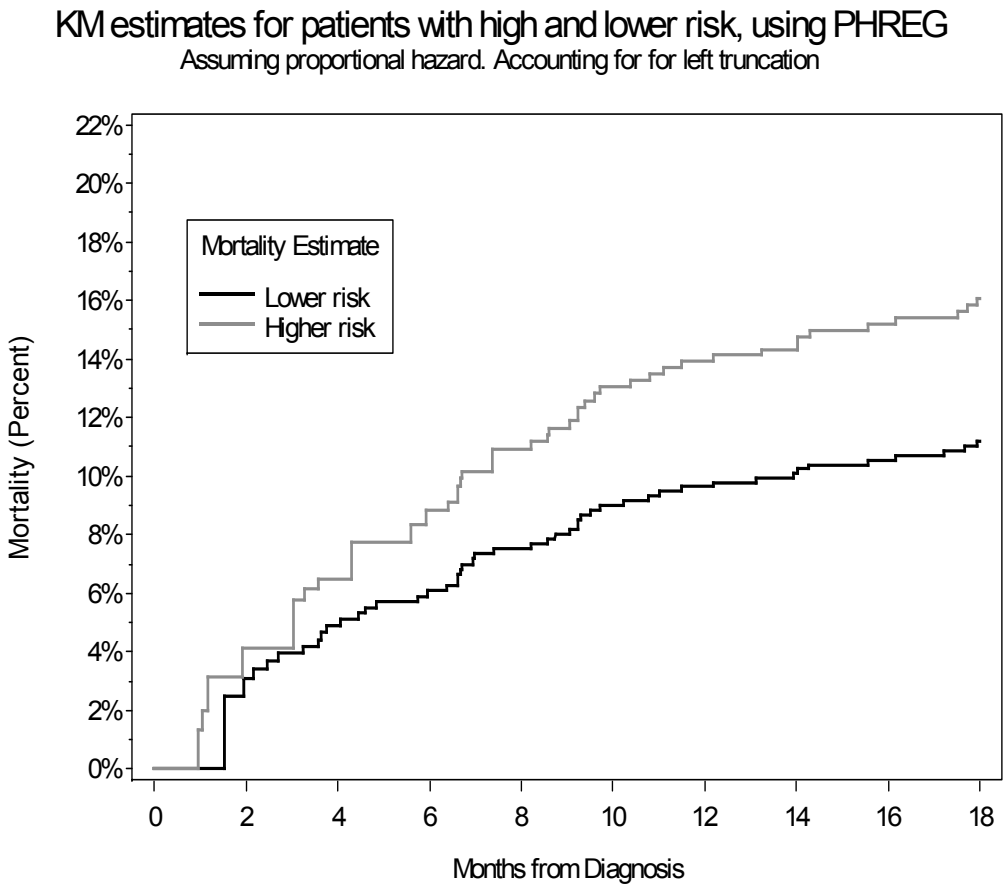
mortality separately for patients with low vs. high test values at the time of diagnosis.

One method for doing this in PROC PHREG is to enter the variable as a predictor in the MODEL statement. In the code that follows, the variable **high_risk** is an indicator variable that has been set to 1 for higher values of the test and 0 for lower values.

```
proc phreg data=temp01;
  model tt_dx_dth*death(0)= high_risk /entry=tt_dx_bl;
  output out=temp07 survival=S;
run;
```

The mortality curves that are generated from this code appear to be nearly parallel (figure 4), and this is by design. PROC PHREG performs regression analysis based on the Cox proportional hazards model, which assumes that the ratio of the hazards is constant over time. The model estimates survival functions for low and high risk that are proportional to each other.

Figure 4



STRATIFIED COX PROPORTIONAL HAZARDS

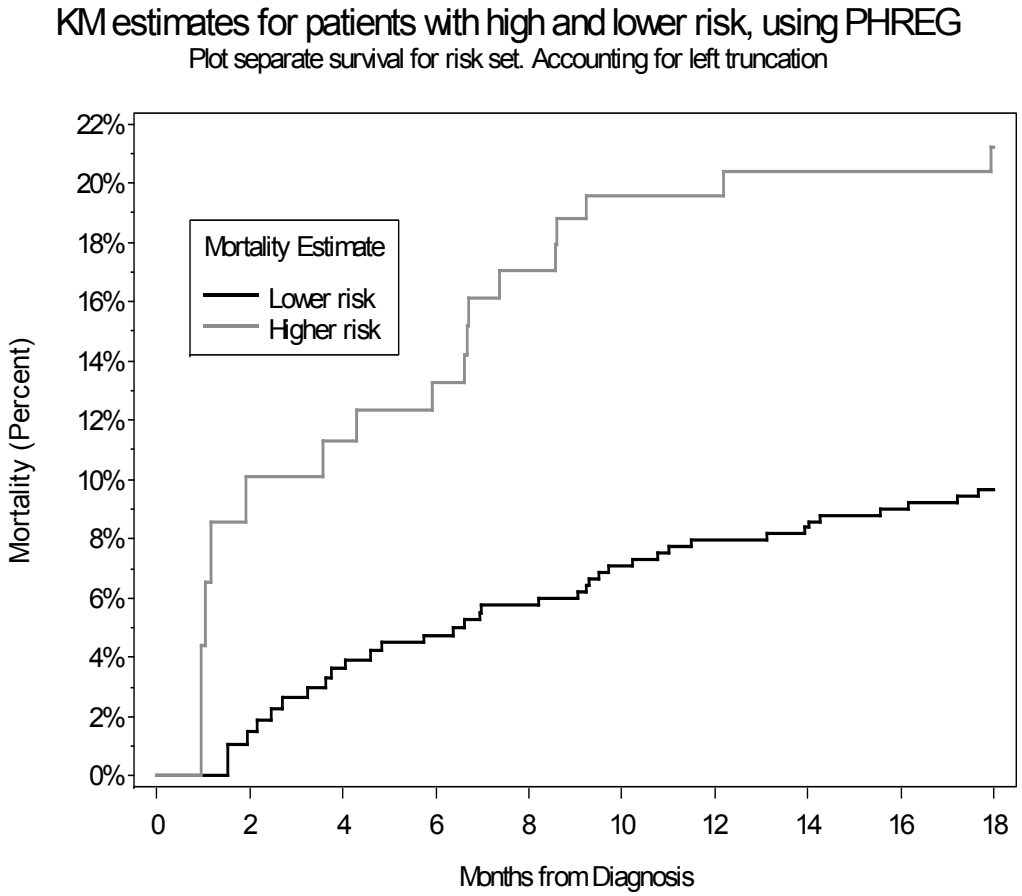
On the other hand, we could allow the underlying hazard function to vary across risk strata. Estimating a stratified PHREG model will allow us to graphically check the proportional hazards assumption. The stratified model consists of a null model with a STRATA statement:

```
proc phreg data=temp01;
  strata high_risk;
  model tt_dx_dth*death(0)= /entry=tt_dx_bl;
  output out=temp08 survival=S;
run;
```

The resulting mortality curves (see figure 5) have a different shape from those where **high_risk** was a modeled predictor. The previously near-parallel lines are much more separated and have varying slopes, suggesting that this

variable has a different effect on the hazard at different times. A comparison of figures 4 and 5 suggests that the proportional hazards assumption is violated, and the stratified model is the better method for generating survival function estimates for group comparisons.

Figure 5



Although we did not perform a log-rank test of the difference in risk levels, the shapes of the curves in figure 5 are consistent with clinical expectations of this factor. For patients with this disease, risk of death is high soon after diagnosis. But over time the high risk patients either die, or they are treated and their test values are lowered.

CODING ALTERNATIVES

The coding methods described above are not the only way to create survival estimates for left truncated data. We will mention a few reasonable alternatives in this section.

COUNTING PROCESS STYLE OF INPUT

The SAS documentation for PROC PHREG identifies two options for dealing with left truncated data: the ENTRY= option described above, and the counting process style of input. It also states that survival function estimates are not computed when you use the counting process formulation, but we found that it is possible to output survival estimates. An email from SAS technical support confirmed that there is a typo in the documentation, which will be corrected in version 9.2.

This method uses a different MODEL statement of the form *MODEL(t1,t2)*censor(list)=variables/options;* where t1 is the entry time, or left truncation time for previously diagnosed patients, and t2 is the event or censoring time. Patients are assumed to be at risk during the entire time from t1 to t2. In order to compute survival function estimates that are equivalent to those produced with the ENTRY= option, it is necessary to use the METHOD=PL option in the OUTPUT statement. The following code produces survival curves that appear identical to those in figure 5:

```
proc phreg data=temp01;
  strata high_risk;
  model (tt_dx_bl, tt_dx_dth)*death(0)= ;
  output out=temp09 survival=S / method=pl;
run;
```

STRATA VS. BY STATEMENT

In this paper we have used the STRATA statement in PROC PHREG for group comparisons. However, we could just as easily have used the BY statement and obtained identical survival estimates. The BY statement does require sorting the data beforehand, which adds an extra step.

BASELINE VS. OUTPUT STATEMENT

In this paper we have used the OUTPUT statement in PROC PHREG to obtain survival function estimates for all observations. The BASELINE statement is an alternative that also outputs survival function estimates, but it outputs estimates only for observations when there is an event. When the model has no covariates, these statements produce identical survival estimates for events, but when there are modeled predictors the estimates are different. In that case the BASELINE statement outputs a data set with survival function estimates at the mean values of the covariates (or at values specified by the COVARIATES= option). The OUTPUT statement, however, creates a data set with survival function estimates at each combination of covariate values. For our group comparison examples, it made more sense to use the OUTPUT statement and therefore we used it throughout.

CONCLUSION

Survival function estimates for left truncated data cannot be obtained from PROC LIFETEST, but they are output easily using PROC PHREG. KM estimators are available from PROC PHREG for any time-to-event data due to the specification of the null model. PROC PHREG accommodates left truncated data with two alternatives: the ENTRY= option for left truncation time in the MODEL statement, or counting process style input. Finally, stratified group survival estimates can be output without having to satisfy the proportional hazards assumption by using the STRATA or BY statement.

REFERENCES

Mayo Clinic Division of Biostatistics. Locally Written SAS Macros. Available at http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm Accessed: April 24, 2008.

SAS Institute Inc., *SAS/STAT® User's Guide*, *Version 9*, Cary, NC: SAS Institute Inc., 2003.

RECOMMENDED READING

Harrell, Frank E. 2001. *Regression Modeling Strategies.* New York: Springer.

Allison, Paul D. 1995. *Survival Analysis Using SAS: A Practical Guide.* Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Aimee J. Foreman
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2119
aimee.foreman@iconplc.com
lsg.iconclinical.com

Ginny P. Lai
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2106
ginny.lai@iconplc.com
lsg.iconclinical.com

Dave P. Miller
ICON Clinical Research
188 Embarcadero, Suite 200
San Francisco, CA 94105
(415) 371-2112
dave.miller@iconplc.com
lsg.iconclinical.com