PharmaSUG2011 - Paper CC16

A Programmer’s Introduction to Survival Analysis Using Kaplan Meier Methods

John Ventre, United Biosource Corporation, Blue Bell, PA
Lisa Fine, United Biosource Corporation, Ann Arbor, MI

ABSTRACT

Survival analysis methods are common in clinical trials and other types of investigation. Programmers are often called upon to program these analyses and produce tables and figures that are often referred to as simply Kaplan Meier Curves. A basic understanding of these methods, their derivation and interpretation of their resulting statistics is provided for a programmer. In addition a presentation for extracting these statistics from PROC LIFETEST is given so the programmer can move from concept to usage.

INTRODUCTION: WHAT ARE SURVIVAL ANALYSIS AND THE KAPLAN-MEIER ANALYSIS?

Survival Analysis represents a set of statistical methods used to estimate lifetime or length of time between two clearly defined events and is sometimes referred to as time to response or time to failure analysis. Survival data is often analyzed in terms of time to an event. For example, in pharmaceutical research, it might be used to analyze the time to responding to a treatment, relapse or death. Analysis of survival tends to estimate the probability of survival as a function of time. For example, estimating the proportion of patients expected to survive a certain amount of time after receiving treatment. Survival Analysis models the underlying distribution of the event time variable (time to death in this example) and can be used to assess the dependence of the event time variable on the independent variables (comparative treatments).

Survival data pose a special case for data analysis in that a study may end or a subject may leave the study before the event is observed. This can be for positive reasons, such as the subject is alive at the end of the study. There are also cases where a subject drops out from a study or a subject is ‘lost to follow-up’ (nowhere to be found) before the study ends. This data cannot be analyzed by simply dropping the incomplete (‘censored’) observations; the fact that the event did not occur is part of the finding. Using the time of censoring as time to the event would also be misleading. In these cases, though the event is not actually observed, the investigator knows that the time to event exceeded a given value (right censoring), such as the last date of contact. A good Survival Analysis method accounts for both censored and uncensored observations.

The Kaplan-Meier curve, also called the Product Limit Estimator is a popular Survival Analysis method that estimates the probability of survival to a given time using proportion of patients who have survived to that time. Kaplan-Meier methods take into account “censored” or incomplete data. Censored observations are incorporated into the analysis up until the time of censoring. The Kaplan Meier analysis makes the assumption that if subjects had been followed beyond the censored time point they would have had the same survival probabilities as those not censored at that time.

DERIVATION OF THE KAPLAN-MEIER CURVE

Suppose a study is performed and a sample of size *n* subjects is collected and for each subject the time to a medical event is recorded (for this example: time to death). You may lose contact with the subjects or they may decide to leave the study. In these cases the subjects are considered censored at the time contact is lost.

In our example, each subject is followed to death or censoring. $t(0) < t(1) < t(2) < ... < t(r)$ are ordered times of death with $t(0)$ is time zero, (failure cannot occur at or prior to start). Depending on the how time is collected more than one subject may die at a given time $t(i)$. You also sample censored times but these are not recorded to the event of interest (death).

First a few assumptions and definitions:

- Let n_j denote the number of individuals alive (at risk) just before time $t(j)$, (n_j includes those who will die at time $t(j)$)
- If an observation is censored at a failure time, $t(j)$, then censoring is assumed to occur immediately after any failures and n_j includes the censored observations. In other words, subjects who happen to be censored at a failure time are considered to be at risk up until and including that time.
- Let d_j denote the number of failures (deaths) at time $t(j)$

For any time t, where $t(k) \leq t < t(k+1)$, the Kaplan-Meier estimate of the survivor function is given by (Kalbfleisch, J. D. and Prentice, R. L. (1980)):

$$\hat{\boldsymbol{S}}(\boldsymbol{t}) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$$

Function is also given as the Product-Limit Method in SAS Institute Inc. (2004).

All this means is…

$$\widehat{\boldsymbol{S}}(\boldsymbol{t}) = \frac{\boldsymbol{n_1 - d_1}}{\boldsymbol{n_1}} * \frac{\boldsymbol{n_2 - d_2}}{\boldsymbol{n_2}} * \ldots * \frac{\boldsymbol{n_k - d_k}}{\boldsymbol{n_k}}.$$

You can see that by moving increasingly across time and multiplying by additional proportions (n_i – d_i)/n_i the survival function will “step” down at an observed failure time.

An example of execution of this algorithm follows in Table A and Table B. Subject (SUBJ) data for a small sample is presented in Table A and is sorted by time (YRS) to failure or censoring. The last column contains an indicator of whether the time represents time to failure (1) or censoring (0).

TABLE A

Subject (SUBJ)	Time (YRS)	Censoring (CEN)
4	3	1
1	6	1
2	8	0
3	12	1
6	12	1
5	21	1

Estimation of the K-M estimate is presented in Table B.

TABLE B

j	*t(j)*	n_j	d_j	(n_j – d_j)/n_j	$\hat{S}(t)$
0	0	6	0	1.0000	1.0000
1	3	6	1	0,8333	0,8333
2	6	5	1	0.8000	0.6667
3	12	3	2	0.3333	0.2222
4	21	1	1	0.0000	0.0000

Taking the *t(j)*=12 as an example of the calculation of $\hat{S}(t)$ the derivation would be:

$$\widehat{\boldsymbol{S}}(\boldsymbol{t}) = \frac{\mathbf{6-1}}{\mathbf{6}} * \frac{\mathbf{5-1}}{\mathbf{5}} * \frac{\mathbf{3-2}}{\mathbf{3}} = \mathbf{0.8333 * 0.8 * 0.3333 = 0.222}$$

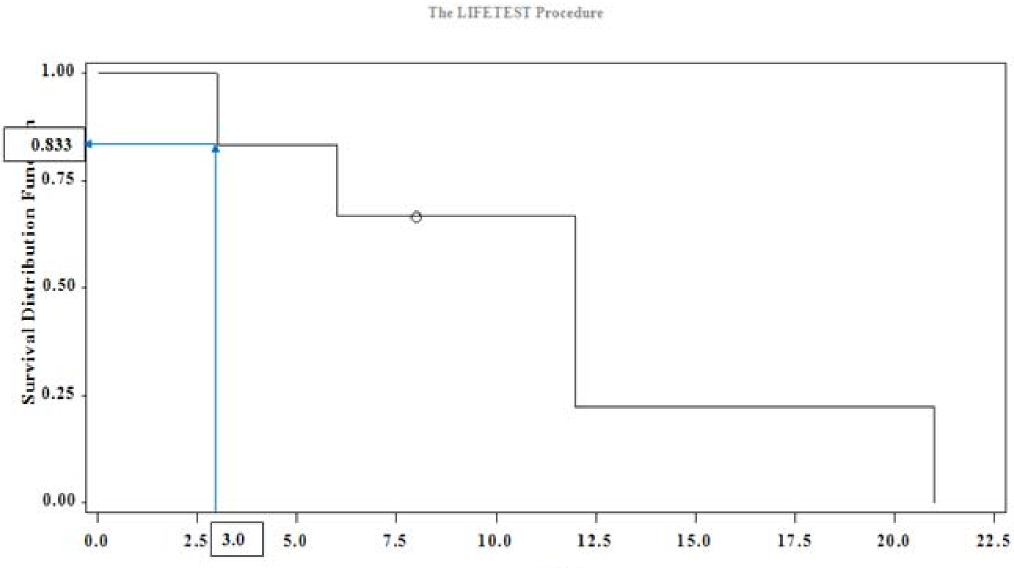
Note that in both Table B and the above calculation the censored time for Subject 2 does not impact the survival function at *t(j)*=8 but does impact survival at *t(j)*=12 by reducing the number of subjects at risk (n_j) by an additional subject.

HOW TO INTERPRET A KAPLAN-MEIER CURVE

Below is an example Kaplan-Meier curve, with the x-axis representing time-to-event in years and y-axis representing probability of survival. This might be an analysis of 6 patients’ time to death following cardiac bypass surgery, with time presented in years. The curve is a plot of the Kaplan–Meier (K-M) estimate of the survival function and is a step function that drops at the time of each event. Thus, the K-M curve presents an estimate of survival as a function of time. Two commonly reported measures in K-M analyses are survival probabilities for either individual time points or time intervals, and the time to event percentiles.

Survival probabilities for any given time can be estimated by projecting up from the time of interest on the x-axis until you hit the survival curve and then moving to the left to the y-axis estimate. For example, according to the figure below the probability of survival at 3 years is 83.3%. (FIGURE 1A)

FIGURE 1A



Time (Years)

The estimated time to event percentiles are easily obtained by starting at the y-axis point of interest (e.g. .50 for 50th percentile), projecting horizontally until the survival curve is met, and heading down to the x-axis time point to get the estimate. It can be seen that the 25th and 50th percentile for FIGURE 1B is 12 years, and the 75th percentile is six years. The 50th percentile interpretation is the estimated median survival time for patients is 12 years.

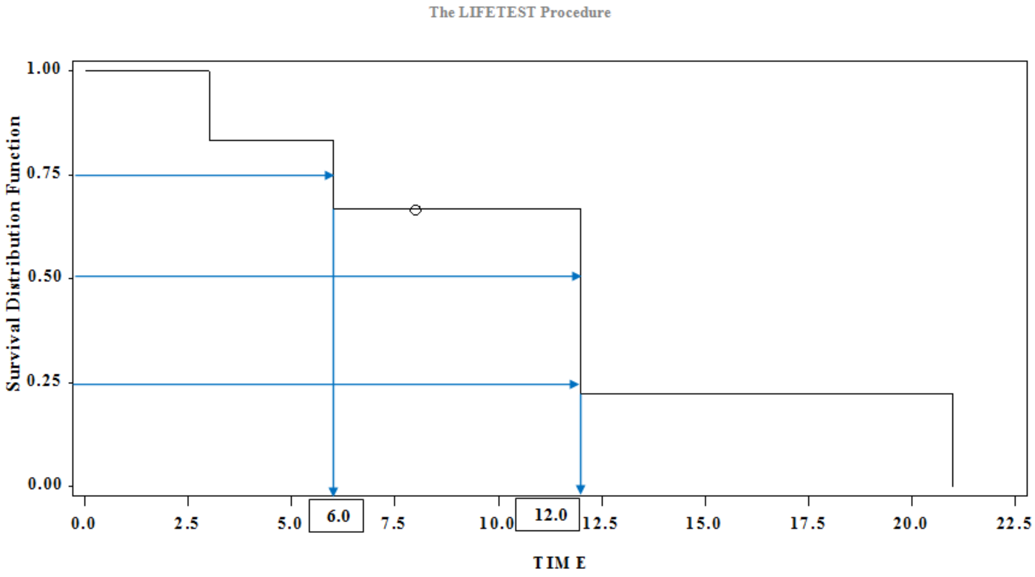


FIGURE 1B

Time (Years)

EXAMPLE KAPLAN-MEIER ANALYSIS

So let's get down to business and do an actual Kaplan-Meier analysis using SAS®. The following section presents syntax, and output for a single sample Kaplan-Meier analysis using the data presented earlier in Table A. Note, that the FIGURE 4 outputs shown are listing output. A later section will demonstrate how to obtain saved Kaplan-Meier estimates so the programmer can reformat the SAS output.

FIGURE 2A displays the general syntax for producing a K-M plot. FIGURE 2B displays the specific syntax for our example. Additional statements and options for generating output files with saved LIFETEST statistics are presented in the next section. These saved statistics will mostly be used for generating tables (versus figures). Note that FIGURE 3A shows the syntax for two different versions of the TIME statements to be used with no censoring and with censoring. You would use only one of these options according to your given scenario, i.e. is there censoring or not?

FIGURE 2A – PROC LIFETEST: KAPLAN MEIER CURVE KEY SYNTAX

```
PROC LIFETEST METHOD=KM PLOTS=SURVIVAL< more options > ;
TIME variable;               **Use this statement for no censoring in your sample (i.e., all events)
TIME variable*censor(list);  **Use this statement with censoring, where “list” are
                               values censor takes on when time is a censored value.
STRATA variable< (list)><...variable<(list)> >;  This is your grouping variable (e.g., Treatment).
RUN;
```

FIGURE 2B - SYNTAX FOR OUR EXAMPLE:

```
proc lifetest method=km plots=(survival);
   time yrs*cen(0);
run;
```

FIGURE 3A below displays survival probabilities. As mentioned earlier this is the listing output. A following section will demonstrate how to save the needed output for later use in RTF tables.

FIGURE 3A THE LIFETEST PROCEDURE

```
                                        Standard   Number   Number
   YRS      Survival(a)   Failure(b)   Error(c)   Failed(d)  Left(e)
  0.0000          1             0           0        0         6
  3.0000     0.8333        0.1667      0.1521        1         5
  6.0000     0.6667        0.3333      0.1925        2         4
 8.0000*          .             .   .                2         3
 12.0000          .             .   .                3         2
 12.0000     0.2222        0.7778      0.1925        4         1
 21.0000          0             1           0        5         0
```

*Censored Observation

INTERPRETATION OF OUTPUT:

FIGURE 3A corresponds to the FIGURE 1A example of how to use a K-M curve.

[a] **Survival** represents the proportion of subjects without an event at the indicated time (YRS) up to but not including the time (row) in the table, in this example, proportion who are alive.

- Note that SURVIVAL probabilities decreased only with the occurrence of an event (and not at an observed censored time designated with an *).

  For example, the proportion of survival went from 1 (or 100%) (no events) at zero time up to 3 years to approximately 83.3 percent at the 3-year point, at which time an event occurred. At 6 years another event reduced the cumulative probability of survival to approximately 66.7%, or worded another way, at 6 years approximately 66.7% of subjects are estimated to be alive. Survival did not change i.e. at 8.000 years, which is a censored observation denoted by the *. Note at 12 years there were two events. Survival is expected to head toward zero, although it will not if the last observation is a censored observation.
- Survival = 1 – the Failure proportion.

[b] **Failure** represents the proportion of subjects with an event (in this example subjects who died).

- As is the case with SURVIVAL, FAILURE changed only with the occurrence of an event. Since FAILURE is the inverse of survival, FAILURE increases with each event.
- Failure = 1 – the Survival proportion.
- Some studies report / graph failure rather than survival.

[c] **Standard Error** represents the standard error of the survival estimate.

[d] **Number Failed** is the cumulative number of events.

[e] **Number Left** (also known as Number at Risk) represents the number of subjects remaining in the study.

Figure 3B below displays probability of event quartiles and corresponds to the FIGURE 1B example of how to use a K-M curve.

FIGURE 3B

```
                 Summary Statistics for Time Variable time
                              Quartile Estimates
                    Point                    95% Confidence Interval
  Percent         Estimate           [Lower                 Upper)
     75            12.0000           12.0000                21.0000
     50            12.0000            6.0000                21.0000
     25             6.0000            3.0000                12.0000
```

Note that the quartiles correspond to FAILURE. For example, 12 years is the time point at which at least 50% of the subjects died. (12 years is also the 75th percentile time point),

HOW TO MAKE THE OUTPUT PRETTY USING SAS

Many SAS users are expected to deliver tables that are formatted beyond the listing style output. The first step to transforming the output is to save the Kaplan-Meier metrics into datasets that can be manipulated or feed other SAS procedures.

STEP 1: OUTPUT PROC LIFETEST DATASETS

SYNTAX EXAMPLE 1 below shows a few of the options available for saving various LIFETEST metrics to datasets (blue font). While it is unlikely you will use all of the outputs created in SYNTAX EXAMPLE 1, the statements are shown for reference. Note, the ‘ods output ProductLimitEstimates = ‘ and the ‘survival out = ‘ statements provide some duplicate, and some unique information. The metrics provided by each statement are shown below so the user can determine which outputs to create for a given report.

SYNTAX EXAMPLE 1

```
ods output CensoredSummary = CENSUM;
ods output Quartiles=QUARTS;
ods output ProductLimitEstimates = KMEST;
proc lifetest method=km plots=(survival);
   time yrs*cen(0);
    survival out = SURVCI conftype=loglog;
run;
```

- **TO OBTAIN SURVIVAL/FAILURE/NUMBER AT RISK;**
  `ODS OUTPUT PRODUCTLIMITESTIMATES = yourdatasetname;`

The statement `ODS OUTPUT PRODUCTLIMITESTIMATES = yourdatasetname;` saves all of the measures seen in FIGURE 3A to a dataset that you name. It also includes a censor variable. Note that the SAS PRODUCTLIMITESTIMATES output assigned censored observations a value of 1 and events a value of 0. This differed from the values specified in our PROC LIFETEST TIME statement (where 0=censor). This is important to keep in mind if you are using this variable later on.

FIGURE 4A.1

```
 PROC PRINT of LIFETEST ProductLimitEstimates Output (KMEST)

 Obs    STRATUM     YRS     Censor    Survival    Failure    StdErr    Failed    Left
 1         1         0        .          1           0          0         0        6
 2         1         3        0        0.8333      0.1667     0.1521      1        5
 3         1         6        0        0.6667      0.3333     0.1925      2        4
 4         1         8        1          .           .          .         2        3
 5         1        12        0          .           .          .         3        2
 6         1        12        0        0.2222      0.7778     0.1925      4        1
 7         1        21        0          0           1          0         5        0
```

The survival out dataset (KMEST) variable names and attributes are shown in the partial PROC CONTENTS output below. Recall yrs was a programmer supplied variable.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
3	Censor	Num	8	1	Censoring Indicator
7	Failed	Num	8	1	Number Failed
5	Failure	Num	8	D8.	
8	Left	Num	8	1	Number Left
1	STRATUM	Num	8		
6	StdErr	Num	8	D8.	Survival Standard Error
4	Survival	Num	8	D8.	
2	YRS	Num	8	8.4	

- ➢ **TO OBTAIN CONFIDENCE LIMITS AROUND THE INDIVIDUAL TIME POINT ESTIMATES:**

```
SURVIVAL OUT = yourdatasetname CONFTYPE=LOGLOG;
```

The **`SURVIVAL OUT = `*`yourdatasetname`*** statement will produce lower (SDF_LCL) and upper (SDF_UCL) confidence LIMITS around survival for each time point. As shown in SYNTAX EXAMPLE 1 the SURVIVAL OUT= statement is part of the PROC LIFETEST code and not an ODS output statement. While **log-log** is the current default confidence limit type it is specified with the **CONFTYPE** option because it is good programming practice in case the default changes. Note, using an OUTSURV= option in the PROC LIFETEST statement would produce confidence limits but without additional options available with the SURVIVAL statement.

FIGURE 4A.2

PROC PRINT of LIFETEST survival out= SURVCI Output (SURVCI)

Obs	YRS	_CENSOR_	SURVIVAL	CONFTYPE	SDF_LCL	SDF_UCL
1	0	.	1		1	1
2	3	0	0.83333	LOGLOG	0.27312	0.97471
3	6	0	0.66667	LOGLOG	0.19462	0.90443
4	8	1	0.66667		.	.
5	12	0	0.22222	LOGLOG	0.00957	0.61472
6	21	0	0	LOGLOG	.	.

The survival out dataset (SURVCI) variable names and attributes are shown in the partial PROC CONTENTS output below. Recall YRS was a programmer supplied variable.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
4	CONFTYPE	Char	8	Transform for Survival Confidence Interval
5	SDF_LCL	Num	8	SDF Lower 95.00% Confidence Limit
6	SDF_UCL	Num	8	SDF Upper 95.00% Confidence Limit
3	SURVIVAL	Num	8	Survival Distribution Function Estimate
1	YRS	Num	8	
2	_CENSOR_	Num	8	Censoring Flag: 0=Failed 1=Censored

- ➢ **TO OBTAIN QUARTILES AND CONFIDENCE LIMITS AROUND THE QUARTILE TIME POINT ESTIMATES:**

```
ODS OUTPUT QUARTILES=yourdatasetname;
```

The statement **`ODS OUTPUT QUARTILES=`*`yourdatasetname;`*** saves all of the measures seen in FIGURE 3B to a dataset that you name.

FIGURE 4B

Proc Print of LIFETEST Quartiles Output (QUARTS)

Obs	STRATUM	Percent	Estimate	LowerLimit	UpperLimit
1	1	75	12	12	21
2	1	50	12	6	21
3	1	25	6	3	12

The quartiles dataset (QUARTS) variable names and attributes are shown in the partial PROC CONTENTS output below.

```
                  Alphabetic List of Variables and Attributes
#    Variable     Type     Len     Format     Label
3    Estimate     Num      8       8.4        Point Estimate
4    LowerLimit   Num      8       8.4        Lower %g% Confidence Limit
2    Percent      Num      8       BEST6.     Percent Label
1    STRATUM      Num      8
5    UpperLimit   Num      8       8.4        Upper %g% Confidence Limit
```

- **TO OBTAIN NUMBER OF EVENT AND CENSORED OBSERVATIONS:**

```
ODS OUTPUT CENSOREDSUMMARY=yourdatasetname;
```

FIGURE 5C

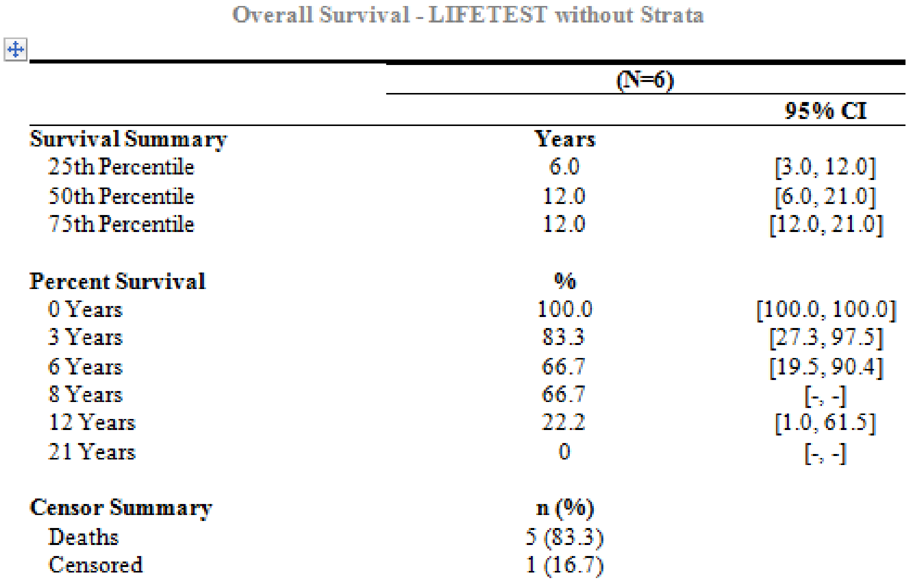
```
PROC PRINT of LIFETEST CENSOREDSUMMARY Output (CENSUM)
Obs        Total      Failed       Censored       PctCens
1            6           5            1            16.67
```

The Censoring Summary dataset (CENSUM) variable names and attributes are shown in the partial PROC CONTENTS output below.

```
                     Alphabetic List of Variables and Attributes
#    Variable           Type           Len          Format     Label
3   Censored            Num             8           BEST8.     Number Censored
2   Failed              Num             8           BEST8.     Number Failed
4   PctCens             Num             8            8.2       Percent Censored
1   Total               Num             8           BEST8.
```

Step 2: COMBINE DATASETS TO OBTAIN NEEDED INFORMATION

After saving your needed LIFETEST output you are able to reuse the data to create formatted reports. For simplicity the goal was to reformat data to look like the table below. For this table we only needed three of the four PROC LIFETEST output data sets shown earlier. The Quartiles ('QUARTS') data set information was used for the Survival Summary (Years). The Kaplan-Meier estimates (Percent Survival) came from the Survival Out= ('SURVCI') dataset. The CensoredSummary ('CENSUM') data set contained the information needed for the for the Censor Summary. After these data sets were combined and labels and numbers were formatted, PROC REPORT was used to produce the table below. Sample code for creating the table is available upon request to the authors.



Overall Survival - LIFETEST without Strata

	(N=6)	
		95% CI
Survival Summary	**Years**	
25th Percentile	6.0	[3.0, 12.0]
50th Percentile	12.0	[6.0, 21.0]
75th Percentile	12.0	[12.0, 21.0]
Percent Survival	**%**	
0 Years	100.0	[100.0, 100.0]
3 Years	83.3	[27.3, 97.5]
6 Years	66.7	[19.5, 90.4]
8 Years	66.7	[-, -]
12 Years	22.2	[1.0, 61.5]
21 Years	0	[-, -]
Censor Summary	**n (%)**	
Deaths	5 (83.3)	
Censored	1 (16.7)	

CONCLUSION

Programmers are often called upon to program and present Kaplan-Meier analyses in Clinical Trials research. The SAS LIFETEST procedure can be used to generate the figures and statistics needed for the analysis. SAS ODS options allow the programmer to extract statistics produced by the procedure and generate a formatted table they require.

REFERENCES

SAS Institute Inc. (2007), Chapter 1 What Survival Analysis Is About, Cary, NC: SAS Institute Inc., Available at http://support.sas.com/publishing/pubcat/chaps/58416.pdf

SAS Institute Inc. (2002-2008), OnlineDoc9.1.3 Language Reference: Dictionary, Fourth Edition, Cary, NC: SAS Institute Inc., Available at http://support.sas.com/onlinedoc/913/docMainpage.jsp

SAS Institute Inc. (2004), SAS/STAT® 9.1 User's Guide, Chapter 40, Cary, NC: SAS Institute Inc.

Kalbfleisch, J. D. and Prentice, R. L. (1980), The Statistical Analysis of Failure Time Data, p12. New York: John Wiley & Sons, Inc.

CONTACT INFORMATION

John Ventre
United Biosource Corporation
920 Harvest Drive, Suite 200
Blue Bell, PA 19422
Work Phone: (215) 390-2242
Fax: (215) 591-2890
E-mail: john.ventre@unitedbiosource.com
Web: www.unitedbiosource.com

Lisa Fine
United Biosource Corporation
2200 Commonwealth Blvd., Suite 100
Ann Arbor, MI 48105
Work Phone: (734) 994-8940 x1616
Fax: (734) 994-8927
E-mail: lisa.fine@unitedbiosource.com
Web: www.unitedbiosource.com