PhUSE 2017

Paper PP08

Surviving your Statistician: A Programmer's Guide to Survival Analysis

James Diserens, Veramed, London, UK

ABSTRACT

Many programmers start their programming career with very little statistical training or understanding. This is not an issue when working with simple statistical methods, such as standard summary statistics. However, programmers providing statistical support alongside statisticians could be are often required to perform more complex analyses one of which is survival analysis. For many non-statisticians, this may be a completely new concept.

This poster explores the basic concepts of survival analysis, aiming to familiarize non-statisticians with the methods and procedures involved in survival statistics. This will allow programmers to gain confidence in the production of survival analysis outputs including Kaplan-Meier and Cox Proportional Hazards analyses, whilst advancing their agility as programmers. The additional benefit of programmers understanding statistical concepts are improved self-checking of their output, which is reflected in reduced comments from the statistician QC'ing.

INTRODUCTION

Programmers working in the pharmaceutical industry can come from a wide range of backgrounds. Many will have experience in sciences and will have knowledge and understanding of more common statistical methods. However very few programmers will have training in some of the more complex methods that are often used in clinical trials.

Working alongside statisticians, statistical programmers can be required to produce or validate statistical outputs. The most effective programmers are those who can understand the purpose of the outputs and the meaning of the results. By having confidence in their understanding of the statistical concepts, the programmer can more accurately produce outputs requested by a statistician and quickly feed back to the statistician any concerns or points of interest. This should reduce the amount of comments from a statistician performing QC and enable them to concentrate on more detailed analysis of the outputs.

A prime example of this is survival analysis. Survival analysis is a concept that many new programmers without a statistical background may not have come across before. This paper will attempt to explain the basic concepts to familiarize a non-statistician with the necessary methods and procedures for survival analysis using manual examples and examples using SAS® code to look at three of the main tools in survival analysis: Kaplan-Meir survival plots, Censoring and Cox-Proportional Hazard (Cox PH) models. This paper aims to help a programmer better survive working alongside a statistician using these methods.

SURVIVAL ANALYSIS

Survival analysis is a commonly used statistical methods in clinical trials, involving the investigation of time-to-event data (the time taken for a specific event to occur or the length of time between two clearly defined events). As the key variable of interest is time, the event measured must be a clearly measurable and pre-defined as a distinct and recordable point at which the subject changes between the two discrete states. Survival analysis can be applied to any outcome that is defined by a clear and measurable endpoint event; such as death, the first relapse of cancer or the failure of a component/device in engineering.

Why is survival analysis used to look at time-to-event data? Why not just use traditional summary statistics or ordinary linear regression? Survival analysis can factor in potential issues that can occur in time-to-event studies and long term clinical trials. Firstly, survival analysis uses censoring which enables data from subjects that did not experience the target event to still be included in analysis. Secondly survival can accommodate the positive skew that often occurs with time-to event data – a larger number of events earlier in the study.

KAPLAN-MEIER

The Kaplan-Meier Estimate product limit is the most widely used method for representing time-to-event data and to estimate the survival function against time. The survival function being the probability of surviving a time period given that the preceding time period has already been survived. In its simplest form, without applying censoring, the survival function is the sum of survival probabilities over small time intervals, calculated as number of subjects who have not experienced the event divided by the number of subjects at the start of the study. The higher the value of the survival function the longer the time for the event to occur.

$$\hat{S}(t) = \prod \frac{n_i - n_e}{n_i}$$

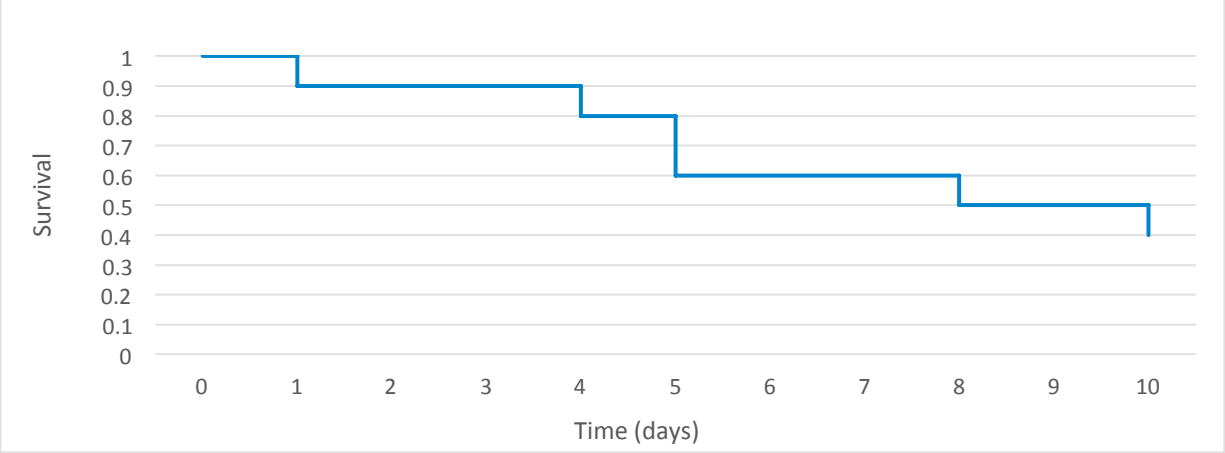
$n_i =$ number of subjects at the start
$n_e =$ number of subjects experiencing event

$S(t) = Survival\ Function$

The Kaplan-Meier Estimate product limit is usually presented as a graph or Kaplan-Meier plot or curve, as a series of steps representing each small time interval. At the end of each step the survival function can be recalculated up to the current time as the sum of all previous time periods, using the equation above. This shows the estimated probability of surviving for a given amount of time, at the end of each step. For the calculation of the survival function it is necessary to assume that the events occurred at the end of the time interval and not as is likely at some point during the interval.

Using the basic example below we can see the relationship between events, time and the survival function and how to plot a Kaplan-Meier curve. Taking 10 subjects reviewed daily for 10 days, and assuming no censoring takes place, as shown in the table below. If after 1 day, a single subject has experienced the event of interest the survival probability for 0 to 1 days is 0.9 and so is the survival function. No event has occurred by day 2 or 3 so the survival function remains the same. By day 4 a further subject has had an event, so the survival probability from day 1 to day 4 is 0.89. The survival function at day 4 then becomes the sum of 0.89 and 0.90. On the following day 2 subjects have experienced an event and the survival function becomes the sum of 0.75, 0.89 and 0.90. This continues until the last day record on day 10. The Kaplan-Meier plot is then plotted as the survival function versus time. Each time an event is recorded the survival function decreases, shown as a vertical line on the plot. Between events occurring the Survival function stays the same and the line stays horizontal.

Time (t)	Number of Subjects Experiencing event	Number of Subjects Remaining	Number of Subjects Remaining	Survival Calculation	Survival Estimate S(t)
0	0	10	10/10=1.0		1.00
1	1	9	9/10=0.90	= 0.90 x 1.00	0.90
4	1	8	8/9=0.89	= 0.89 x 0.90	0.80
5	2	6	6/8=0.75	= 0.75 x 0.80	0.60
8	1	5	5/6=0.83	= 0.83 x 0.60	0.50
10	1	4	4/5=0.80	= 0.80 x 0.50	0.40



CENSORING

Censoring is a key part of time to event data and survival analysis, although not unique to it. Sometimes the exact lifetime of a subject cannot be recorded, because the event being studied could not be observed before the end of the trial. These subjects still provide useful information and should not be ignored as removing them from analysis could introduce a bias into the study data. Censoring allows the subject to still be included in the analysis, for the duration that was observed. There are three different types of censoring; right, left and interval. Most common in clinical trials, is right censoring.

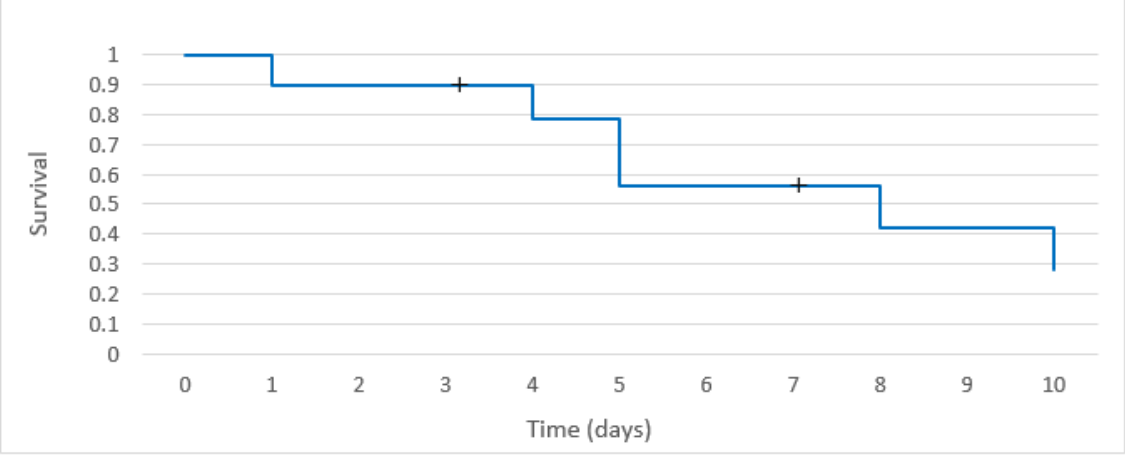
Right censoring is performed when the start time of the subject is recorded but the time taken for an event to occur is unknown. Often with real world data in clinical trials, there could several reasons why a subject is censored: A study ends before the subject could experience the event, the subject withdrew from the study or was 'lost to follow-up' or the subject has another event that stops the event of interest from happening (e.g. dying before experiencing a different event of interest).

This example above shows a simple example and the result is obvious without the use of survival analysis. When working with data where subjects are censored, this process becomes more useful. Assuming the same example except this time two subjects withdraw part way through the study, one on day 3 and one on day 7.

The survival function is the same up until day 4. Now during the period from day 1 to day 4 we only consider the 8 subjects who completed up to day 4, and the survival probability for that period becomes 0.875. Again, between day 5 and day 8 a subject withdrew so instead of using 5 subjects we only calculate the survival from the four subjects with complete data. Although we know only have 2 out of the 10 subjects surviving until day 10 the Survival Function is now 0.281 because subjects where time-to-event is unknown are factored in. As seen on the plot below, on

Time (t)	Number of Subjects Experiencing event	Subjects Censored	Number of Subjects Remaining	Survival Probability	Survival Calculation	Survival Estimate S(t)
0	0	0	10	10/10=1.0		1.000
1	1	0	9	9/10=0.90	= 0.90 x 1.00	0.900
4	1	1	7	7/8=0.875	= 0.875 x 0.900	0.788
5	2	0	5	5/7=0.714	= 0.714 x 0.788	0.563
8	1	1	3	3/4=0.750	= 0.750 x 0.563	0.422
10	1	0	2	2/3=0.667	= 0.667 x 0.422	0.281

Kaplan-Meier plots points of censoring are marked, typically with a +, but the survival function is not effected.



EXAMPLE IN SAS®

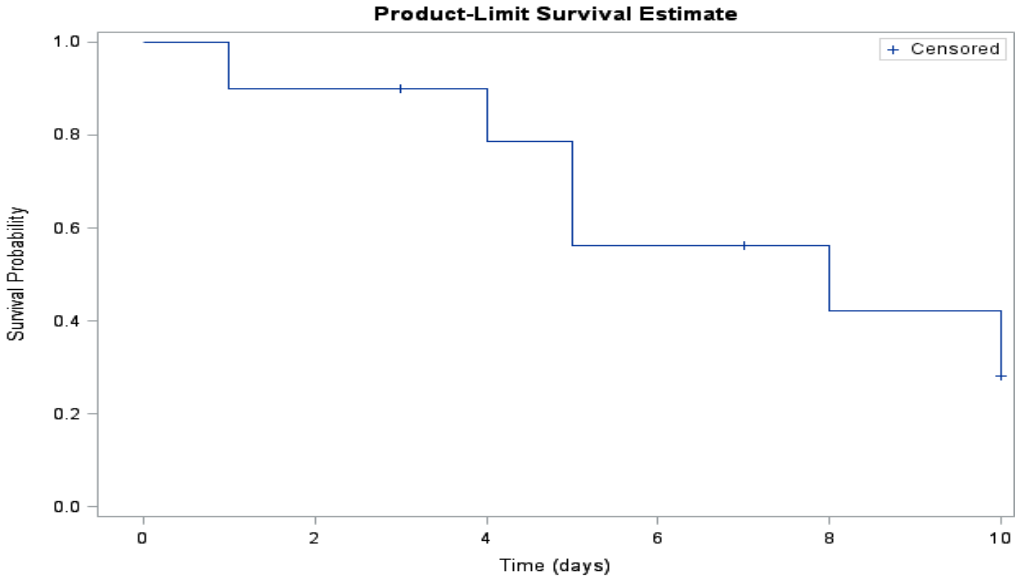
Kaplan-Meier plots can be produced in SAS® by using the PROC LIFETEST procedure. The code below uses the input dataset *ex1*. The input dataset should be subject level with subject ID [Subject], time to failure [time] and a numeric flag to indicate if censoring has occurred [cnsr]. The code for the example above will then look like this:

```
proc lifetest data=ex1 outsurv=km1 plots=s;
   time time*cnsr(1);
run;
```

The OUTSURV= option produces a dataset containing the survival estimates, censoring and confidence intervals on each day that an event was recorded. The PLOTS=S option instructs SAS to display a plot of the estimated survival. The time statement is required for PROC LIFETEST, and defines the variable containing the failure time – in this example called time. This can then be followed by an * with the censoring variable – cnsr – and the reference variable to right censor the estimate. In this example the variable cnsr is used with a value of 1 to indicate that the subject was censored and 0 to show the event occurred.

The outputs of this example are shown below and we can see how this matches the manual example that we performed above.

	Time (days)	Censoring Flag: 0=Failed 1=Censored	Survival Distribution Function Estimate	SDF Lower 95.00% Confidence Limit	SDF Upper 95.00% Confidence Limit
1	0	.	1	1	1
2	1	0	0.9	0.4730092714	0.9852813934
3	4	0	0.7875	0.3808815232	0.9425909522
4	5	0	0.5625	0.2094212437	0.8091709215
5	8	0	0.421875	0.111028773	0.7125672743
6	10	0	0.28125	0.044741403	0.595754959
7	10	1	0.28125	.	.



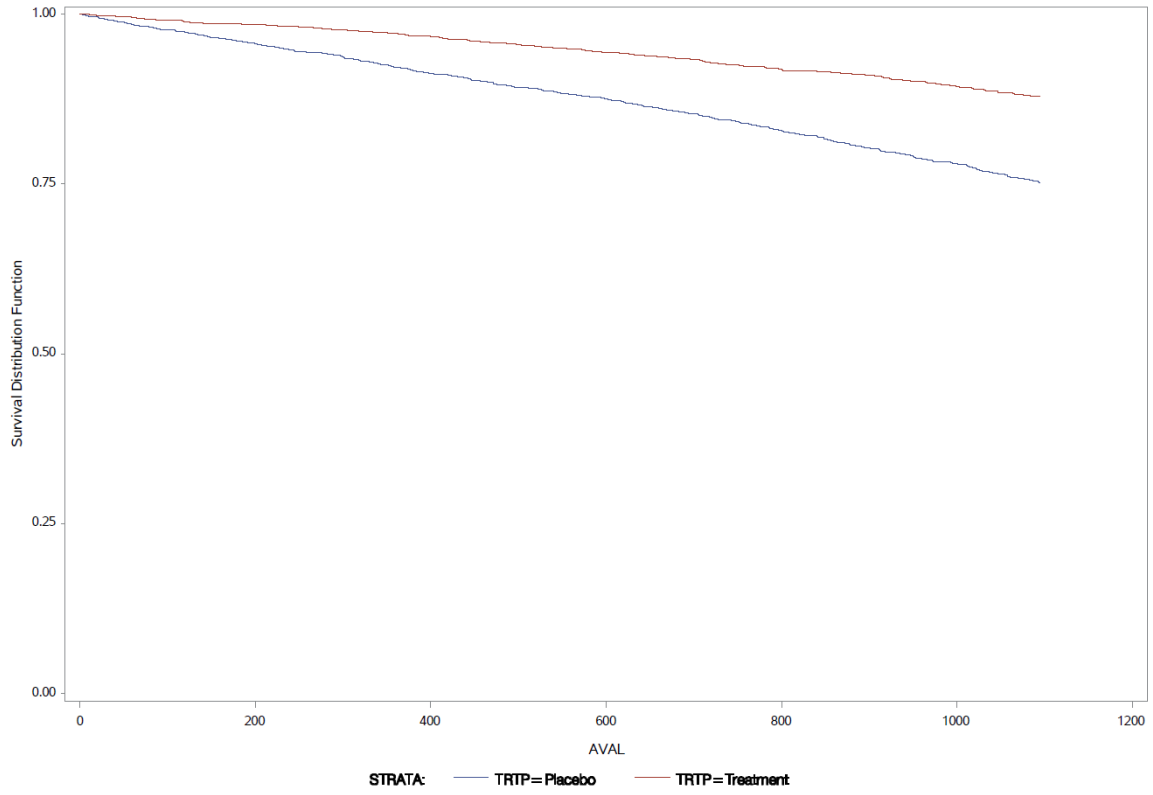
Now let's consider a larger example, more to the scale of a clinical trial, and using multiple treatment arms. Using a dummy study between a treatment and placebo, lasting 3 years and consisting of 8000 subjects between the ages of 40 and 80 where the primary endpoint is time to death. This data is likely to be collected in the ADTTE AdaM dataset. An example of the important variables is shown for 5 subjects below.

STUDYID	SUBJID	SEX	AGE	TRTP	TRTP	PARAMCD	PARAM	AVAL	CNSR
STY110	301000	M	76	2	Treatment	T2DTH	Time to Death	1020	1
STY110	301001	M	67	2	Treatment	T2DTH	Time to Death	395	1
STY110	301002	F	79	1	Placebo	T2DTH	Time to Death	1029	0
STY110	301004	M	76	1	Placebo	T2DTH	Time to Death	1095	1
STY110	301005	M	58	1	Placebo	T2DTH	Time to Death	1095	1

For the purpose of clinical trials, we now need to split the analysis by treatment group to compare the effect of treatment versus placebo. To do this with PROC LIFETEST, we use the `strata` statement. With many subjects being investigated there are a lot of censored subjects and the markers can obscure the line. Therefore, we introduce the `NOCENSOR` option on the plot to keep the lines clear.

```
proc lifetest data=adtte outsurv=outsurv plots=s(nocensor);
    time aval*cnsr(1);
    strata trtp;
run;
```

The plot below shows the initial output and comparison between survival for the Treatment and Placebo arms. With the increased number of subjects, the lines become smoother and can be considered as a continuous curve. The curves show a higher survival for subjects on Treatment over Placebo for 3 years. This suggests that there is longer time-to-death for these subjects, not that there is a lower chance of subjects receiving the treatment dying. In this example, it appears that subjects in the placebo group have approximately an 80% chance of surviving beyond 3 years and the Treatment group has approximately 90% chance of surviving beyond 3 years. However, we can improve the plot further to make this easier to see.



The plots option on PROC LIFETEST, has its limitations, so it is often better to use the outputted survival tables and PROC SGPLOT to customize the display. Using PROC LIFETEST in combination with `ods productlimitestimates` statement, as shown in the code below, gives a dataset containing the Kaplan-Meier Estimates and the constituent parts (surv1 – shown for 5 observations below). Using this output and the output from PROC LIFETEST containing the confidence intervals we can build the plot manually.

```
ods listing close;
ods output productlimitestimates=surv1;

proc lifetest data=adtte outsurv=outsurv cs=none;
    time aval*cnsr(1);
    strata trtp;
run;

ods output close;
ods listing;
```

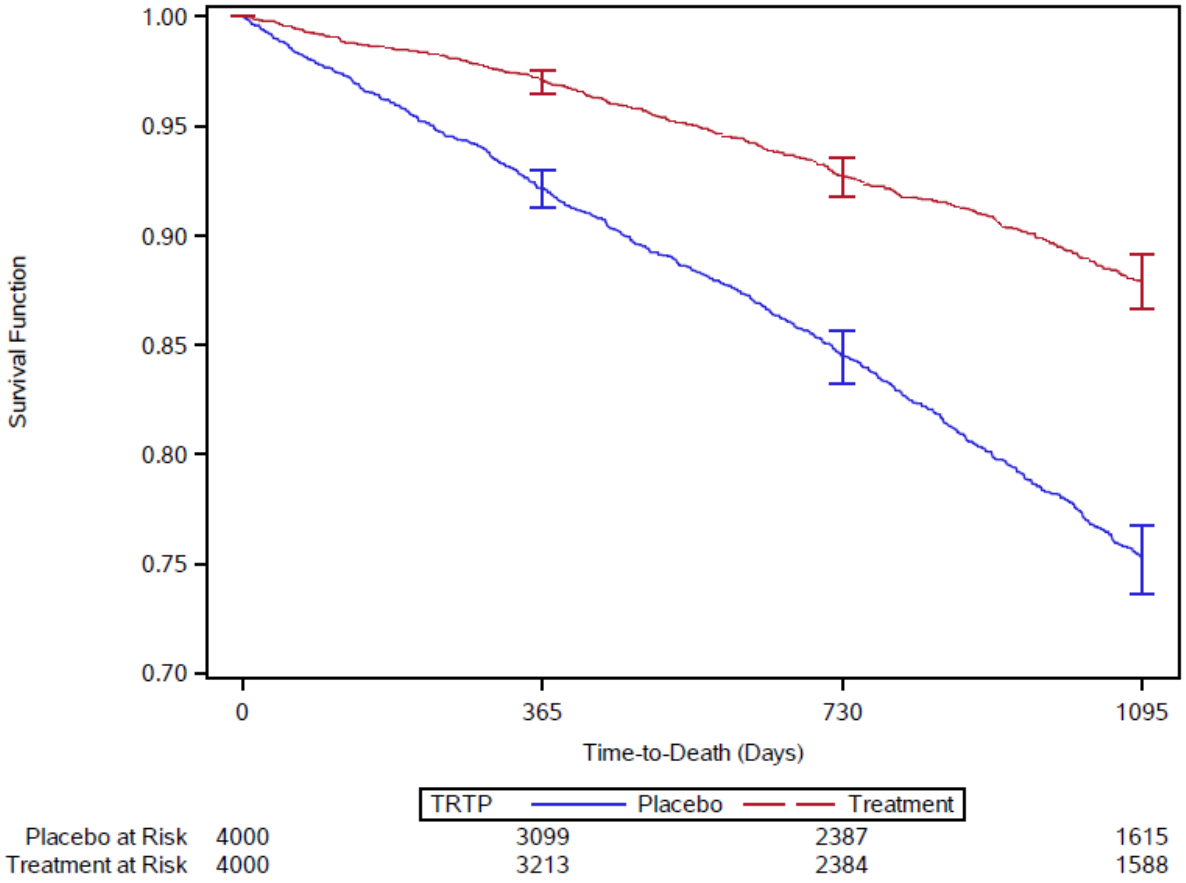
	STR
1	
2	
3	
4	
5	

4000
4000

The dataset o*utsurv* will contain the confidence intervals. Merging this will surv1 and creating the variables lower and upper containing the confidence intervals at the desired time points we have the dataset surv2 ready to create the Kaplan-Meier plot. Now using a PROC SGPLOT series we can manually plot the Kalan-Meier as aval vs survival, as shown below. From there we can start to make changes to improve the plot. To make the graph easier to interpret, we focus on the area of interest on the y-axis (0.7 to 1.0) and pick relevant time intervals on the x-axis (every year). Adding confidence intervals helps make comparisons between the placebo and treatment groups. Having confidence intervals at every step would obscure the lines, so in this example we will add them at each year. By using the variables low_ci and high_ci that are populated only at the desired time points we can add confidence intervals. It is also helpful to include the number of subjects at risk for each group over time. For PROC SGPLOT we have done this by using an `axistable` statement, again using yearly intervals, with variables that are only populated at the desired times.

```
proc sgplot data=surv2;
  series x=aval y=survival / group=trtp;
  scatter x=aval y=survival / group=trtp yerrorlower=lower yerrorupper=upper
  xaxis label='Time-to-Death (Days)' min=0 values=(0 to 1095 by 365) max=1095;
  yaxis label='Survival Function' values=(0.7 to 1 by 0.05);
  xaxistable plarisk trtrisk;
run;
```

The plot used in the example above is drawn as the survival function, decreasing from one to zero. Kaplan-Meier plots can also be made to show the percentage of subjects experiencing events and these go from 0 to 100 (or 0 to 1 if using probabilities). Instead of seeing the survival rate of the population we now have a plot showing the event or hazard rate.



ASSUMPTIONS AND INTERPREATION

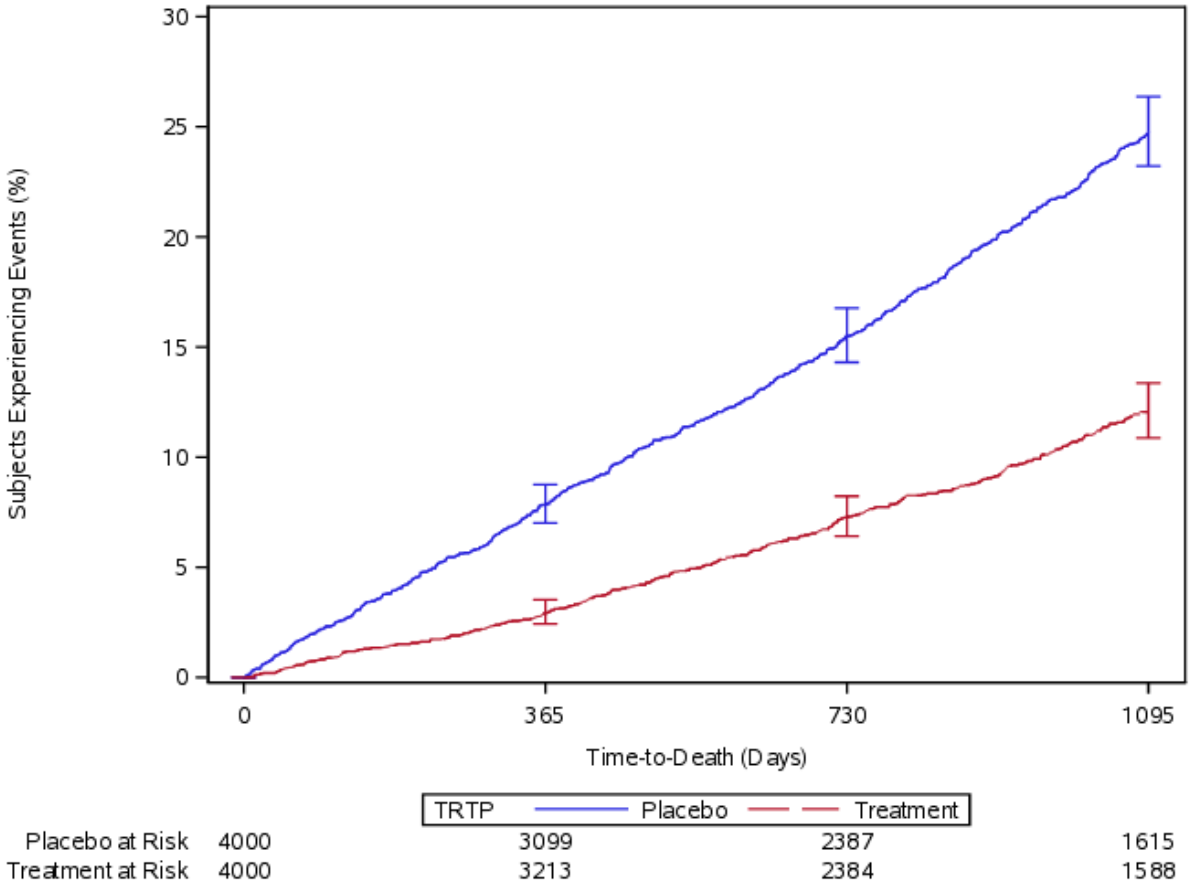
Having produced a Kapan-Meier plot it is important to know how to properly interpret the results. If we look at the plot of the Kaplan-Meier estimate above, it would be tempting to conclude that subjects receiving the treatment are less likely to die than those receiving placebo, but this is not necessarily true. All we can say is that subjects who received treatment took longer, on average, to experience the event than those receiving the placebo.

The results of a Kaplan-Meier plot can be communicated in various ways: Proportion surviving at a specific time point: approximately 75% of patients in the placebo group and 90% of patients in the treatment group have survived for 3 years (1095 days). Median survival, the time taken for half of the subjects in the group to experience the event. As fewer than 50% of subjects do not experience the event within 3 years, we cannot calculate this in the example.

When looking at the results of the Kaplan-Meier plot it is important to remember several things:

- Assumes censored subjects have the same probability of survival as those who continue.
- Before any subjects are censored, the estimate is an exact representation of the data.
- More frequent follow-up of subjects increases accuracy of estimate.

As with any other statistical procedure it is important to be aware of the limitations and potential influences for the results. Due to censoring the left-hand side of the graph will have greater precision as the calculations are being performed on a greater sample size. If the follow-up time is long enough to allow most of the subjects to either experience the event or to be censored, or if subsets with small populations are used, then the right-hand side of the graph will be difficult to interpret. As the sample size becomes low, the standard error rises. With low sample sizes, a single event will result in a large step on the plot.



COX PROPORTIONAL HAZARD

Cox proportional hazards (Cox PH) regression is a method for investigating the effect of variables on time to event data, using survival times and censoring. When the assumptions of Cox regression are met, it can provide a better estimate of survival probabilities and hazard rate than with Kaplan-Meier. Cox PH allows us to compare the effect of different covariates on survival times using the Hazard Ratio.

The Hazard Ratio is used to show the difference in time to event between groups, i.e. the relationship between an exploratory variable to a reference value. It is the ratio of the hazards of the two different groups on the time-to-event, where hazard is the instantaneous probability that an individual would experience an event, given that this individual has survived to that particular point of time without experiencing the event. Cox PH models assume that hazard functions for two different levels of a covariate are proportional for all different time points.

The hazard ratio is equivalent to the ratio of probability of an event between the two groups. Therefore, assuming the placebo group is the reference group, then the following can be said:

- If the Hazard Ratio = 0.5: then half as many patients in the treatment group are experiencing an event compared to the control group.

- If the Hazard Ratio = 1: then the event rates are the same in the two groups
- If the Hazard Ratio = 2: then twice as many patients in the treatment group are experiencing an event compared to the control group.

It is often preferable to consider the reduction in risk in a group compared to the another. This can be calculated as

- Reduction in risk = 1 – HR (usually converted to %).

EXAMPLE IN SAS®

In SAS, the PROC PHREG procedure is used to perform Cox PH analysis.

In the model statement, the variable aval (Time in days from the ADaM dataset adtte) is crossed with cnsr (The censoring variables). The value in brackets indicates the value of cnsr for subjects that have been censored. The variable after the equal sign, trtp, tells SAS to calculate the ratios between.

```
proc phreg data=adtte;
   class trtp(ref='Placebo');
   model aval*cnsr(1) = trtp / risklimits;
   hazardratio trtp;
run;
```

The results will be shown in SAS: Output as below under 'Analysis of Maximum Likelihood Estimates'. This shows, amongst other things, the hazard ratio, the p-value and the 95% confidence intervals for the treatment group compared to the placebo group.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
TRTP	Treatment	1	-0.81763	0.06542	156.2225	<.0001	0.441	0.388	0.502

Using the example from above, without Covariates, comparing Treatment against the reference group of Placebo, gives a hazard ratio of 0.441 (0.388, 0.502). This means that there is a 56% reduction in the risk of dying for subjects in the treatment group compared to the placebo group. With a p-value <0.0001 this can be considered significant.

	Hazard Ratio (95% CI)	**Reduction in risk of Death**	**p-value**
Treatment vs. Placebo	0.441 (0.388, 0.502)	55.9% (49.8%, 61.2%)	**<.0001**

COVARIATES

As with linear regression modelling, Cox PH models allows the use of covariates. A covariate is a variable that can be observed that may affect the outcome of interest, but is not the main variable of interest (which will often be study treatment). By including covariates into a model the estimate of the treatment effect can be shown after taking into account the effects of these other variables. The covariates used in the model should be decided upon before analysis takes place. These should be chosen input from experienced clinicians and statisticians and included in the SAP.

```
proc phreg data=adtte;
    class sex trtp(ref='Placebo');
    model aval*cnsr(1)=trtp sex age / risklimits;
    hazardratio trtp;
```

```
run;
```

Again the results will be shown in SAS: Output as below under 'Analysis of Maximum Likelihood Estimates'. This time as well as having results for the overall treatment group, there are values for gender and each age.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
TRTP	Treatment	1	-0.82980	0.06565	159.7548	<.0001	0.436	0.383	0.496	TRTP Treatment
SEX	F	1	-0.05201	0.06454	0.6494	0.4203	0.949	0.837	1.077	SEX F
AGE	40	1	0.45988	0.29348	2.4554	0.1171	1.584	0.891	2.815	AGE 40
AGE	41	1	-0.19931	0.38683	0.2655	0.6064	0.819	0.384	1.749	AGE 41
AGE	42	1	0.45944	0.29761	2.3832	0.1226	1.583	0.884	2.837	AGE 42

In the example here we have two covariates: Age and Sex. However, this data comes from a randomised clinical trial, so we would expect these to be randomly distributed across all arms, so we not expect to see much change in the hazard ratios. The reduction in risk is again 56% for subjects in the treatment group compared to the placebo group.

	Hazard Ratio (95% CI)	**Reduction in risk of Death**	**p-value**
Treatment vs. Placebo	0.436 (0.383, 0.496)	56.4% (50.4%, 61.7%)	<.0001

CONCLUSION

The best statistical programmers are those who have a good understanding of the statistical concepts that they are working with to respond to the request of a statistician. This enables the programmer to not only accurately produce the requested outputs but to also feed back to the statisticians any concerns or points of interest in the data. By understanding the processes working with time-to-event data and the manual calculations used in survival analysis the production the programmer should be better prepared to spot anything unusual in their outputs. This will enable them to focus on any interesting areas within the results and spot any potential errors when performing self-QC.

CONTACT INFORMATION

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

James Diserens
Veramed Ltd
5th Floor Regal House
70 London Road
Twickenham / TW1 3QS
Work Phone: +44 (0)20 3696 7240
Fax:
Email: james.diserens@veramed.co.uk
Web: www.veramed.co.uk